Leveraging digital forensics and data exploration to understand the creative work of a filmmaker: a case study of Stephen Dwoskin's digital archive

Leveraging digital forensics and data exploration to understand the creative work of a filmmaker: a case study of Stephen Dwoskin's digital archive

Zoe Bartliff[1], Yunhyong Kim[1], Frank Hopfgartner[2], Guy Baxter[3]

[1] Information Studies, School of Humanities, University of Glasgow, Glasgow, UK
[2] Information School, University of Sheffield, Sheffield, UK
[3] Museum of English Rural Life (MERL), University of Reading, Reading, UK

Abstract

This paper aims to establish digital forensics and data exploration as a methodology for supporting archival practice and research into a filmmaker's creative processes. We approach this by exploring the digital legacy hard drives of the late artist Stephen Dwoskin (1939-2012), who is recognised as an influential filmmaker at the forefront of the shift from analogue to digital film production. The research findings of this case study show that digital forensics is effective in extracting a timeline of hard drive activities, data that can be explored to reveal clues about the artist's personal/professional history, stages of creative processes, and technical environment. The paper further demonstrates how this is related to current thinking around user-centred archival workflow and understanding of creative processes. The broader impact of the work for advancing digital archiving and research into creative processes is highlighted, concluding with a discussion of how, going forward, the approach can be coupled with deeper content analysis to reveal what influences editing choices taking place over time.

Keywords: digital forensics; data exploration; creative process; filmmaking; timeline analysis; digital archives

1. Introduction

1.1 Research context

In 2013, the University of Reading (UoR) acquired the archive of the experimental filmmaker Stephen Dwoskin[1], presenting a unique opportunity to approach his work, through one of the most comprehensive holdings of a single artist filmmaker in the UK[2]. The research presented in this paper is part of the research project, "The Legacies of Stephen Dwoskin's Personal Cinema",[3] developed with the broad objective to understand what the Dwoskin archive can reveal about the aesthetic, cultural, and historical conditions, the experience of disability, independent creative practice, and economic and technological changes which independent filmmaking underwent from 1960's onwards.

Apart from his role in the independent filmmaker scene of the 1960s and 1970s, Dwoskin's struggle with disabilities and how this influenced his social circumstances and artistic practice holds pertinent interest in the

[1] University of Reading Special Collections MS5502
[2] Stephen Dwoskin was best known for his central role in establishing "the London Filmmakers' Co-op (LFMC) and then the Independent Filmmakers' Association in the 1970s (an organisation that paved the way for Channel 4). The earlier work of LFMC, in turn, contributed to the founding of LUX and LUX Scotland active international arts agencies that promote artists' moving image practices.
[3] AHRC grant AH/R007012/1

interpretation of his films. He contracted polio at a young age which increasingly affected him throughout his life, and his "films never hid his disability; indeed, the human body – his own or other people's, in all imaginable states of pleasure and pain – became his central subject"[4]. He adopted technical solutions to develop this subjective and personal approach to audio-visual production, distinct from structural narrative cinema. He experimented with a number of techniques to manipulate the scene, for example, to "distort, slow and stretch the image" to convey "embodied affects such as claustrophobia, abjection and desire" [36]. This might suggest media formats and technology (both analogue and digital) as essential components in understanding his artistic development process.

Stephen Dwoskin's archive includes correspondence, working materials, unpublished writings, and photographs. In addition to this, it also includes twenty hard disk drives suspected to contain unfinished versions of films, email correspondence, personal documents, his unpublished autobiography, as well as other materials that might shed light on Dwoskin's personal experiences as an artist with disabilities. The hard drives have the potential to be a rich academic resource for exploring the diverse content, media formats and technologies embedded within an artist's working environment and creative processes. The research presented here illuminates digital forensics and data exploration as an effective foundation for unearthing the story behind the data to: 1) provide a summary of hard drive content and associated technical environment to inform archival workflow and strategy, and, 2) capture information relevant to working patterns highlighting stages of artistic development. The present paper will use six of the Dwoskin hard drives (those successfully acquired as complete forensic images at the time of writing) as a case study to demonstrate how digital forensics and data exploration can be used for these purposes.

1.2 Motivation for exploring the Dwoskin archive

Dwoskin's hard drive collection distinguishes itself in a number of ways that make it of special interest to the broader archiving community and to digital scholarship research. It is the only collection of hard drives naturally accumulated by a single artist over more than a decade during the course of their life. The archive, comprises approximately 12 TB,[5] and, is possibly the largest personal digital archive to be discussed in the published literature (e.g. [82] explores less than 17GB, [23] estimates a potential archive of 500GB comprising the activity of more than one individual but does not actually maintain the archive). Whist this presents a unique opportunity to test the scalability of forensics tools for archival purposes, the computational and human labour needed for data/content management and review poses a challenge.

This is exacerbated when it is understood that the context in which Dwoskin's drives were used and the material that they contain is yet to be fully understood. It is common practice for institutions, when taking on a personal digital archive, to ensure a degree of pre-acquisition intervention as a part of their workflow (e.g., [84]), for example, to include a discussion with the creator regarding the content and usage, and to identify necessary provision for the preservation of the content (e.g. technical requirements and permissions/consent). However, such preparation is not always possible ([70]), as is the case with the Dwoskin digital archive, which was received after his death. The drives, therefore, could be considered a fossilised artefact with content representing Dwoskin's per-

[4] https://www.theguardian.com/film/2012/jul/12/stephen-dwoskin

[5] Collective size of all the drives. A breakdown of the content size for each of the six drives examined in this study is given in section 3.2.

sonal/professional practices, something which could be an advantage for preserving their original context of creation. On the other hand, the consequence is that the disk type, the file structure and the content of the drives is not necessarily ideal for archival review or collection. With little to no supporting documentation with regards to the contents, the researcher/archivist would be entering examination of the large volume of data blindly, which could be alleviated with creative data exploration techniques.

The age of the drives, some stretching back to the late 20th century, also suggests a variety of challenges associated with hardware, software, and format obsolescence, a well-recognised obstacle in digital preservation (cf. [40]). However, for the same reason, it represents a valuable resource for exploring digital scholarship in the humanities (cf. [89], [38]), and, as of yet, seems to be unrivalled in scope, as a compendium of information communication and audio-visual technology in the context of a single independent filmmaker. From the initial analysis, some of which is presented in this paper, the content of the drives comprises diverse complex multimedia, typically associated with audio-visual production, and their edits, over a long period of time. This raises concerns regarding the "fidelity of the migration process" that might take place to allow accessibility at the archive (cf. [40]). At the same time, the range of time period covered by the hard drives opens up the opportunity for a longitudinal study of creative process recorded in its natural environment without the employment of artificially imposed instruments for recording data such as workshops (cf. [34]) or structured personal information management guidance ([84][60]). The content available in video editing tools, for example, records edits, and opens up the opportunity to examine fine-grained editing choices (cf. [64]) as well as allowing an insight into the relationships between these choices and other information on the drive being accessed concurrently.

The content is further distinguished by the extensive range of personal, private and/or confidential information within (e.g. emails, financial documents). This suggests looming legal and ethical obstacles for wide scale access to the archive, and could impose a substantial demand on archival review processes especially given the size of the hard drives. Conversely, the availability of personal communications could help reveal conflicts and collaborations that were instrumental in the artist's life and artistic development and his contribution to the independent film landscape.

1.3 Going forward

The challenges and opportunities outlined in Section 1.2 highlight the value of this case study for archival practices and the opportunity for researching an artist's creative process. It also suggests directions regarding what we would like to capture as the summary of hard drive content and patterns of the artist's working practice.

For archival purposes, the size of the collection demands an automated approach to generate a summary comprising, for example, a list of files, a timeline of when they were created and/or modified, and where they are located across the drives. To support mitigating issues related to obsolescence, it would also be useful to understand the technical environment of the content, such as file formats and associated software. File types (for example, video, text, image, audio) can also help formulate strategies for reviewing the content for personal, private or confidential information. Understanding the artist's creative process also demands similar information, although in this latter case, the timeline of file activity might become accentuated to determine broader relationships between files and their environments couched in development activities over time.

In this paper, we demonstrate that digital forensics can be used to create the necessary timeline of development activities over time. Digital forensics, to be discussed further in Section 2.1 and 2.4, is routinely used for the

recovery and examination of digital evidence in criminal investigations ([22], [11], [21] [98]). It encompasses a wide set of methodologies which are employed in the process of recovering, preserving, validating and analysing digital evidence in its original form and context, most commonly for the purposes of criminal investigations. Here, combining it with data exploration, we hope to repurpose it for uncovering the narratives associated with creative work from digital evidence.

We propose digital forensics and data exploration as a useful bridge for balancing creative research and archival practice, toward user-centred archives. This will be discussed further in Section 2, when we elaborate on the research objectives of this paper. This is followed by a description of our data workflow in Section 3, and our findings in Section 4. Implications of the findings will be discussed in Section 5, concluding with a discussion regarding conclusions and future directions of this research in Section 6.

2. Research objectives and related work

In Section 2.5, we will break down our research objectives further to show how we will approach extracting the timeline as discussed in Section 1.3. First, however, in Section 2.1 and Section 2.2, we will demonstrate how our research aligns with current thinking around archives and their connection to creative work. We will follow with a definition of *creative processes* in Section 2.3, and, finally, we will review relevant approaches to timeline generation and summarisation, to clarify how our research sits within the general literature, and, to explain why the forensic methodology is appropriate for current research (Section 2.4).

2.1 The archive and the end-user

The issue of user interaction and the user's needs has come to the fore of archival studies since the 1960's ([15]) and it became apparent that there was a divide in the manner that archivists approached their work. One branch, comprising the traditionalist 'keepers of records' style archivists, was primarily content focused which contrasted with those archivists who were 'use-centred' ([26][42][59][49]). Whilst it is evident that there is a use to the methods of the former, particularly in the preservation of records and their context, the latter is equally as important if those preserved records are to be accessed and made useful in the modern world. The more theoretical aspects of this debate have tailed off in recent years and the focus has moved towards the practical application of user-friendly archival structures. With the increasing trend towards digitisation and the push for open access research within academia,[6] the need for a user-focused approach to archival studies has become undeniable ([107][88][85][59][26]). It has been shown, for example, that historians are most often considered as the primary users of archives ([69]) and that the traditionalist content focused archives can be upheld as exclusive and so only open to those who already access them on a regular basis ([59]) and who are conducting 'serious', rather than 'frivolous' research ([88]).

Beyond this, it has become evident that traditional approaches to archival formation and management are not suited to the vast quantities of data that archives are now required to process, but that they may in fact prevent the users from obtaining the data that they are seeking ([85], [86]). This situation is ultimately undermining one of

[6] https://www.ukri.org/funding/information-for-award-holders/open-access/ref-2021-open-access-policy/

the key functions of the archive, and is destined to become a more complex issue as the already vast quantities of digital material increase requiring additional metadata, improved interfaces and a more digitally focused skillset for the archivists and the users ([104][83][93]). The Dwoskin archive, both digital and analogue, as already mentioned above, is the first of an incoming trend. Currently unique in scholarly literature not just for its content as would be expected of any archive, but also for the scope of that content. Spanning Dwoskin's work until his death at 73, and containing numerous content types, the Dwoskin archive is of potential interest to a huge variety of researchers, from historians to artists, casual enthusiasts to digital forensics and multimedia specialists. Such a broad scope of potential users of the archive, not to mention the variable disciplinary backgrounds and skillsets of those researchers, adds demand on making sure that the digital content of the drives is readily accessible and in a usable format.

The use of digital forensics for curatorial purposes in the context of personal archives has been explored in a number of places (e.g. [75],[57],[106],[82]). The potential of forensic disk imaging, as part of the archival workflow, to unpack content bundled into storage devices has been long accepted (e.g. see [82]). Similarly, in terms of data extraction, John notes the changing landscape of "personal and cultural information", and presents data exploration methods such as "mining concepts and relations" as approaches "effective in identifying entities, events and associations of interest in unstructured textual content, and in stimulating systematic hypotheses" (see [57]). This assertion is supported by numerous other studies, particularly as regards text extraction and analysis ([78]). These previous studies, however, are very much driven to meet archival needs (e.g. for example, to support preservation workflows) rather than the needs of the end-user. [82] and [62], for instance, emphasise digital forensics to support data authenticity, reliability, and integrity, and/or to recover data content for accessibility, and metadata collection tools such as DROID[7] to profile content file technical metadata (e.g. file format, size – cf. [44], [105]), selected to meet the fields of metadata and cataloguing standards developed for preservation, such as PREMIS[8]. Likewise, aims often revolve around supporting "chain of custody" (see [106]), and the preservation of the digital artefacts perceived to facilitate experiencing the final creative products, for example, in new media art (e.g. [24], [29], [62]).

The research in this paper differs from previous works by steering the focus of digital forensics and data exploration away from characterising content files to activities surrounding those files (e.g. creation, modification, access; software/hardware usage) that might summarise the original context of creation[9], or to the needs of end-users studying such contexts. These end-users make more tangible the archive's "designated community", a concept that has been rather elusive and, by being elusive, has been less than useful in developing preservation strategies (e.g. [12],[61]), since its inception in the Open Archival Information System (OAIS) reference model ([66],[61]). In our case study, this community comprises those engaged in and/or researching creative practice and associated influences in the context of moving images and sound. By focusing on access of the archive dataset at both a detailed and a general level, whilst also preserving both the original context of the content and its original

[7] The UK National Archive ongoing PRONOM file formats registry and file format identification tool 'DROID' employs a number of techniques, including an examination of the file signature and extension to provide an accurate identification of a file, so long as that file type exists within its database. (http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/)

[8] https://www.loc.gov/standards/premis/

[9] Terry Cook, Public Lecture, "Landscapes of the Past: Archivists, Historians and the Fight for Memory" http://www.culturaydeporte.gob.es/dam/jcr:c15b65cd-851c-4a81-9528-6ef87984e417/conferencia-terry-cook.pdf

form,[10] the proposed digital archival structure minimises the archivist bias inherently present in material selection ([19], [43],[73]) and enables a wide spectrum of potential research.

2.2 Personal archives and creative work

As artists are increasingly moving into a digital environment for their work, there is the increasing likelihood that the personal archives pertaining to their work will be digital in nature [63]. Several studies highlight the significance of personal digital archives for the analysis of an artist's work, for example regarding the importance of teaching individuals, such as musicians [7]. In response, a wealth of literature has emerged critiquing the digital preservation strategies already employed by those charged with their management ([96]). The general consensus is that, whilst people tend to be aware that preservation is an issue, the uptake of effective strategies is variable. On the whole, where it is utilised, replication of content is the primary approach ([96],[70], e.g., backing up on a device, storage on social media websites, email) and that this is due, at least in part, to in-built technological functions. However, at least within the given studies, this was the sole activity and there was very little administration that went along with these practices, particularly with regards to an organised filing system or the deletion of outdated items [96]. There is rarely employment of a formal filing system, at least without the presence of certain 'triggering events' such as reaching the capacity of available storage or acquisition of new equipment ([96]). Whilst personal digital information management is important for digital preservation, to an extent it could mask some of the more intricate details of an individual's creative practice, particularly with regards to the timeline of creation, administration of projects and their own personal associations between files. The artificial intervention suggested by, for example [7], whilst admittedly preserving content, interferes with the naturally accumulated process metadata. In this paper, we reveal groupings of an individual's creative activities and transitions between stages of creative process, by focusing on bringing to the forefront and analysing this metadata.

2.3 Defining creative process

The purpose of the current paper is not to get entangled in the dispute over how we might comprehensively define the stages of a creative process. Instead, in our study, we will adopt the most common theme across the research area and demonstrate how our data exploration methodology might serve to reflect these stages. The research around creative processes has an extensive history (e.g. [5], [103], [102], [80], [47], [45], [100], [96], [79], [17], [41], [71], [34], [91], [54],[97]). The conversation around the complex concept of creative process is on-going (e.g.[91],[65]) across a number of fields, for example, in relation to education (e.g. [102], [92]), to influences of filter bubbles (e.g. [8]), to advertisements (e.g. [101]) and to visual art (e.g. [34]). Lubart gives a thorough review of the evolution of thoughts stemming from the four-stage model to their variants in the 20th century (see [71]). A recurring point of reference across the literature is a four-stage model representing 1) preparation, 2) incubation, 3) illumination, 4) verification, which is intended to capture the core phases underlying

[10] Without archivist intervention except the necessary redactions to conform to the legal requirements of data security given the above mentioned presence of private and/or confidential documents/data present on the drives.

a creative process. The four-stage model has been attributed to Wallas (e.g. [102],[103]), but may have taken its current form through Guildford's interpretation of Wallas's model (see [45]). The four stages are presented below as described by [65]:

- **Stage 1 (Preparation):** Knowledge items relevant to the problem are formed. Associations start to arise between them and they are manipulated in small, overlapping groups representing different views but are not yet organized.
- **Stage 2 (Incubation):** Attention is moved away from the problem's network. This phase models the relaxation of constraints and thus, new associations are intuitively formed. At this stage, the network is not stable and contradicting knowledge items tend to compensate each other.
- **Stage 3 (Illumination):** A potentially stable network of knowledge items starts to form that grows rapidly after a moment of sudden illumination.
- **Stage 4 (Verification):** Follows a moment of sudden illumination and the user starts to focus again on the problem, reflect on the current stable network and breakthrough happens whereby the user finds a way to connect previous and current ideas together.

One strong theme reflected in determining the four stages of creative process is related to tracking 'knowledge items' (Stages 1 & 3) as they are formed, finalised and revisited. For this, it is essential to understand the timeline of the activities evidenced on each of the hard drives. Another strong theme hinges on periods of 'illumination' and 'rapid growth' (Stages 3 & 4), likely captured in patterns of varying intensity of the artist's activity, for example, seasonality. Finally, to understand 'associations' and 'networks' (Stages 1, 2 & 4), clusters of concurrent or collocated activity become central. It would be of interest to see how the activities, intense periods of activities, concurrency, collocation relate to the professional/personal events in the artist's life (e.g. filmography), and, to the creative material/environment (e.g. file types, software used).

It seems intuitive that "the affordances of particular systems, environments and technologies are often integral to creative processes" ([28]). For example, when it comes to analogue creative works such as paintings the use of advanced imaging and analytical techniques are a viable way of allowing researchers to "see the creation process" ([67], [1]). In his book, 'Tracking Changes', Kirschenbaum [64], similarly, presents an argument, through the anecdotal exploration of the working practices of numerous authors, that the tools utilised in the process of creating written work mediate between the inspiration and skill of the creator and the physical output that is the result of it. The characteristics of the tools, therefore, can contain within them the imprint of the creative process and the context surrounding it. In a traditional text, such analyses could involve an investigation of analogue media – the paper, the ink, and the paint – and, equally, any evidence of an iterative process – edits, glosses, and discarded versions. Kirschenbaum presents (p26, 209-10, 220,[64]) numerous narratives in which scholars have made ground-breaking discoveries through exploration of such information. He draws a parallel of analogue media to the digital through a review of multiple authors' "word processing" practices on the computer, but then equally comments that it is unknown whether similar discoveries would even be possible with the migration of the creative process to a digital and therefore transitive environment.

A number of artificial intervention methods have been used to study creative processes (e.g. participant observation; diaries; brain imaging, games - [35], [13], [34], [50], [53], [65]). These, however, often involve

artificial intervention to unearth relevant processes (e.g. an assigned task; incorporation of activities or tools into the workflow to allow monitoring the process; use of equipment to monitor context specific indicators such as eye movement or brain activity), and they tend to confine the study to a relatively narrow temporal scope. The framework does not normally allow the visualisation of influences in the span of a person's life history (cf. [81]) nor does it allow studying creative processes in its natural environment. As we will see in Section 2.4, this situation could potentially change when we have access to the personal digital devices where they did their creative work. In this case, through the use of digital forensic methods, we argue that, not only is it possible to conduct traditional forensic examinations ([31]) to establish a timeline of events surrounding or leading up to a particular occurrence, but also to examine the particular patterns of activity that reveal the preparation, incubation and verification patterns of Dwoskin's creative process.

2.4 Timeline generation and summarisation

2.4.1 Timeline analysis and digital forensics

Timeline analysis forms a standard approach used for criminal investigations in digital forensics, where an investigator would test an hypotheses by establishing "what happened when (sequencing)" in relation to "who interacted with whom (linkage)", the "origin of a particular item (evaluation of source)" and "who was responsible (attribution)" for activities on the disk drive [21]. This research is more concerned with the context of the data and the reconstruction of a specific event or series of events and, on the whole, is utilised for criminal investigations and/or the prediction of aberrant behaviour within a population or dataset. As such it is usually combined with content analysis (e.g. identification of different media types, recurring patterns/keywords) examined at different levels of the computer system [20], toward an understanding of, amongst other things, activities associated with the working patterns of different users, clusters of materials that are accessed at the same time and recent activities related to software.

Digital forensics tools, in addition to helping to access legacy media which are inaccessible through other means, are invaluable in reconstructing events from such devices using metadata evidence ([21] [52]) to illuminate, amongst other things, file changes over time. File metadata associated with activities on a person's computer (e.g. timestamps associated to creation, modification and access and corresponding filenames and size) can be as valuable a source of information for understanding a person's interests as the content of the files themselves ([14] [48] [52]). Security experts, for example, have long recognised that metadata such as the timing, length, and frequency of access to information can reveal a substantial amount of information on what and who we are interested in, and what might be important to us, at any given time ([95]). For example, in reference to [72], researchers showed that phone metadata alone[11] could be used to determine whether, for example, somebody was a heart attack victim, an owner of a semiautomatic weapon, a home marijuana grower, and/or someone who had an abortion. In fact, recent work demonstrates how metadata can be applied to reconstruct events to explain the creation of data [58].

[11] Historical call and text message [Short Message Service (SMS)] metadata from device logs volunteered by people. No additional information out with the device used for inference. Facebook data used for ground truth.

Similarly, the metadata on a filmmaker's computer has the potential to reveal context surrounding their creative process by revealing those things of interest and of importance to them at a given point in time. This kind of metadata is documentation about the digital environment in which the artist worked, and this kind of documentation "can sometimes communicate more about a work and how it is experienced than its physical manifestation" or any notion of completed work ([28]). In this paper we aim to employ timeline analysis anchored to such metadata embedded in the system to show how this can be used to summarise a filmmaker's working pattern over time, especially when coupled with file/folder name and extension analysis, and when discussing the first two stages of creative process (preparation and incubation – see Section 2.3).

The social, material, and temporal dimensions of analysis involved in digital forensics aligns well with the notion of "distributed creativity" as discussed by Glăveanu [39] and immediately reframes the digital forensics methodology for criminal investigations, as a new potential methodology for research into creative practices. For example, the same forensic investigation procedures could illuminate editing processes, patterns of creative flow, and technological choices. These insights, in turn, could guide choices about further sampling, coding and interpretations in relation to deeper content analysis, for example, to interrogate the interplay of "clichés", "stereotypes" and moments of "inspirations" in the filmmaker's work. The methodology for reconstruction of events can go beyond critical research to support "synthesis" (cf. [16]) as Dwoskin's legacy is inherited by future filmmakers and through doing so enrich the archival record for the potential user .This paper sets out and describes the initial findings from the first stages of exploring this methodological approach.

2.4.2 Content-driven timeline extraction

Cross-document timeline extraction related to a selected event and/or topic has been studied in a number of researches already (e.g. [77]) and narrative extraction received recent attention in special issues (e.g. [2]) where some research brought timeline summarisation together with narrative generation (e.g. [9]). All these works share the similar objectives to extract events related to target topic, extract the temporal expressions and the temporal relationships, and carry out co-reference resolution ([77]). In some cases, the focus is on extracting temporal features (e.g. [108]; [55]), in other cases, it is more focused on tracking the topic and/or entity (e.g.[3][4]). A large quantity of literature revolves around the automation and refinement of existing methodologies in response to the increasing volume of data in our day to day lives ([90][74]). For example, [32] notes how existing natural language processing and information retrieval tools are insufficient for handling large quantities of data in real time, before evaluating some of the existing methods of text summarisation. [6] and [4] both have a focus on the need for good visualisation of timelines, the former through the utilisation of graphs and the latter through a specialised, multi-layered interactive timeline tool. The utilisation of traditional flat timeline models fails to account for the multifaceted and multi-themed nature of event-based timelines and simply prohibits the necessary depth of information required to present the larger story. Social media has received a noticeable amount of attention within scholarly literature, in particular with regards to the issues of finding the behavioural trends that could inform on group opinions and actions outside of the social media application. Similarly, such techniques can be utilised with, for instance, news reports ([76]) in order to find trends and interrelated patterns of behaviour with the intention of modelling and, to an extent, predicting behaviour ([18]). Increasingly, however, there has been research into timeline and content analysis of other media types, particularly video material (e.g. [94], [68]).

The works discussed in this section aim to extract information from a high-volume corpus, usually in real time, and so construct a cogent and condensed narrative from that information ([74][86]). This approach usually has an internal focus, with the primary aim of providing a narrative for the evolution of a specific dataset and interpolating how that data was altered from one data point to another. This could play an essential role in the examination of deeper creative processes, for example, related to specific changes in media content that signal moments of inspiration or influence, most notably for understanding the transition between stages 3 and 4 (see Section 2.2). In the work presented in this paper, we are interested in what we can learn about transitions between stages 1 and 2, although, content-driven timelines do form a key part of the research planned for the future of the project.

2.5 Focused research questions

In this paper, we present data exploration aided by timelines generated by a digital forensics approach (introduced in Section 2.3) as a means of creating an ordered narrative summary of the filmmaker's artistic development practice – a summary relevant to digital archives (Section 2.1 and 2.2) and their community of researchers investigating creative processes.

The research presented here contributes to the discourse by utilising Stephen Dwoskin's digital legacy as a case study. to explore the following research questions:

- **Personal and/or professional history**: what is the extent to which Stephen Dwoskin's personal and/or professional history can be reconstructed from our dataset using digital forensics? Is it possible to identify periods of activity likely to be linked to Dwoskin's filmography and, further, to isolate evidence that his work might have been affected by seasonal variations?
- **Stages of creative process:** is it possible to identify stages of Dwoskin's creative process through timeline analysis? More specifically can we relate the emerging profile of Dwoskin's work habits to the lifecycle of files, from their creation, to the final point of modification and then to the most recent instance of access, which might reflect creative process?
- **Materials and Environment:** is it possible to summarise, from the activity timeline, the broad technical medium/environment in which Dwoskin worked, for example, the digital formats and/or software that Dwoskin utilised over time? More specifically does file type clustering or file path exploration provide a clue to his environment?

Although these research questions are centred around the artist's creative processes, they also address the archival requirements outlined in the introduction. For example, the first research question can only be answered by summarising the list of files contained in the collection, their distribution across the different drives and their relationship to each other in terms of activities on the hard drives. Likewise, the second research question, by looking at the file's lifecycle, identifies dates of creation relevant to content files, and the third research question provides a comprehensive summary of the technical environment, flagging up possible software required or migration processes that might mitigate issues of obsolescence. In fact, by putting it into the context of questions of creative process pertinent to the designated community, the archive would be able to reveal additional information about how these changed over time. For example, the archive would be able to demonstrate how files

and environment related to each other in relation to selected events in the life of the artist (e.g. film production), an advantage of user centred archive design.

3. Methodology

3.1 Data workflow

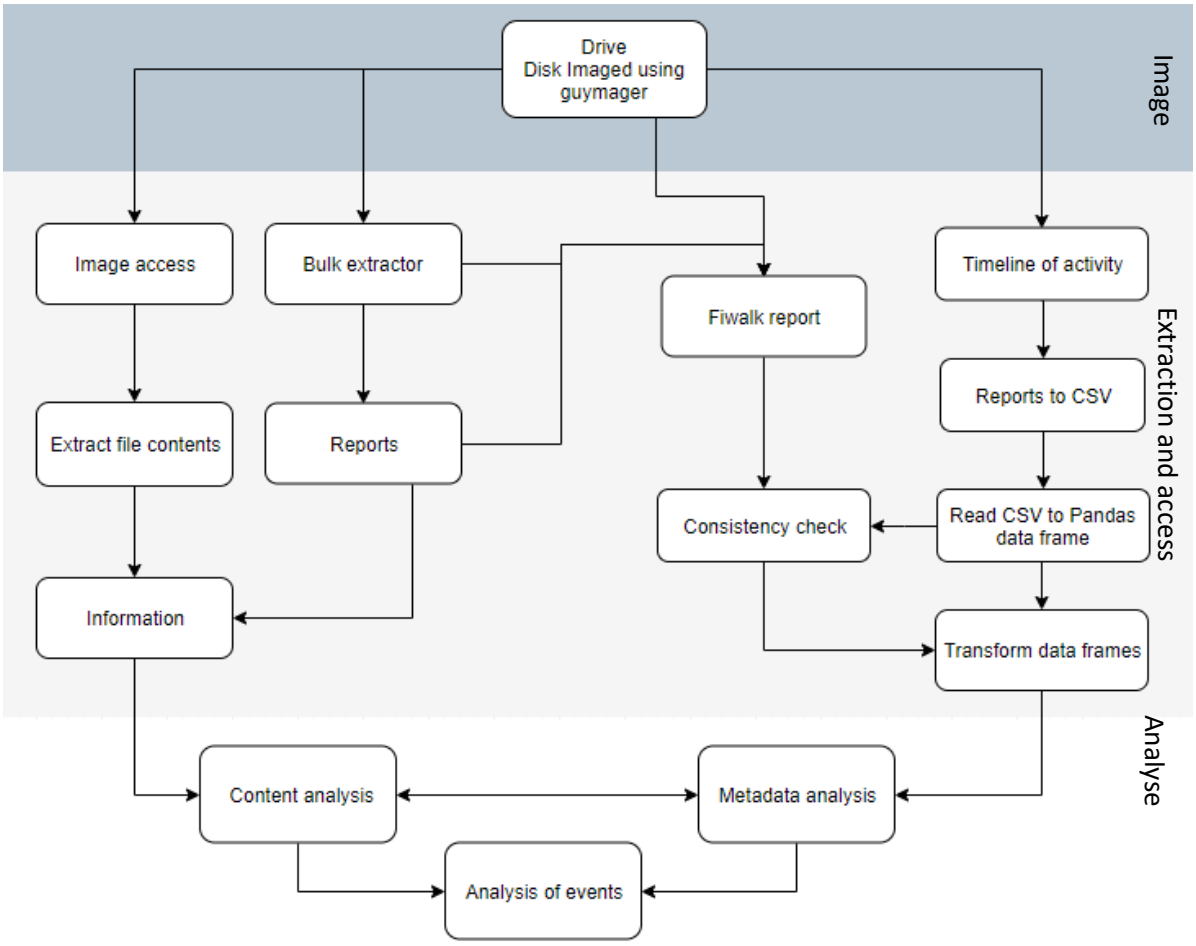

Figure 1 - Workflow for hard disk drive analysis: showing the process from imaging to content and metadata analysis

The data workflow proposed here can be divided into three interdependent parts (as demarcated by the colour gradation in Figure 1): forensic disk image creation (top), data extraction and access (middle), and metadata and/or content analysis (bottom). The role that each of these play in analysing patterns of activity, possible clusters of materials, and technologies, for example, is explained in more detail in the following subsections.

3.2 Disk image creation

Reliable metadata and content analysis is only possible after the hard disk drive is captured as a forensic disk image, a bit-by-bit, sector-by-sector direct copy of a physical storage device. This is crucial for several reasons:

- It ensures that disk information (e.g. temporal metadata) is not inadvertently changed during analysis.
- By performing an original disk image and storing the original disk, it enables exact reproduction of analysis methods on the original evidence.
- The method makes it potentially possible to capture information invisible to the operating system (e.g. deleted files).

The research here is carried out on images acquired using Guymager on the BitCurator platform[12] using the EnCase forensic image format. As a record for the archive, basic reports can be generated using bulk extractor [37] and fiwalk (integrated into The Sleuthkit[13]). Bulk extractor extracts content features usually considered to be useful for forensics (e.g. email addresses, urls, and jpegs) while fiwalk extracts file system metadata and links the features extracted by bulk extractor to files and associated timelines, where possible. Together they present an overview of disk composition prior to any deeper content analysis.

Each of Dwoskin's twenty hard disk drives have been assigned a label from 01 to 20 by the holding archive to aid in the process of cataloguing. In the absence of obvious clues as to the contents of the drives, these numbers were assigned without design or intent. The six drives examined in our study were selected through convenience sampling in as much as they are the first six drives successfully imaged using Guymager. In Table 1, we have presented the size of each of the six drives, the number of content files reported, and total content. While the images might still contain files that had been deleted, neither The Sleuthkit nor bulk extractor was able to identify deleted files.

Disk	Storage size (GB)	Imaged size (GB)	No. of files (fiwalk report)	Content (GB)
01	320	79.5	9993	104.29
02	1000	859	report failed[14]	747.55
03	500	380	2015	357.53
04	250	83.2	3469	101.54
06	250	198	1075	233.26
10	500	317	1094	320.48

Table 1 - Overview of hard drives in the study. Sizes are in gigabytes (GB).

3.3 Timeline generation

To understand Dwoskin's working pattern, we propose the exploration of the timestamps found on the drives, which can be used to build a timeline of activities. Each point in time can be annotated by the number of actions taking place summarising the disk drive usage over time. The Sleuthkit is able to acquire metadata relevant to generating timeline of file activities. An example of The Sleuthkit extraction is shown in Table 2. Each entry in

[12] The BitCurator packages open source digital forensics tools in a Linux operating system to enable their "incorporation into the workflow of archives/library ingest and collection management environments, and provision of public access to the data".

[13] https://www.sleuthkit.org/

[14] Both the bulkextractor and fiwalk reports failed for 02. There are many potential reasons for this and they are still under investigation

the timeline (see Table 2) can be further analysed to differentiate the type of actions that are associated to files (column Type in Table2): the last date associated with creation (b), modification (m), access (a), and record change (c), which could, depending on the file, all be the same date or all different. This MACb time analysis is a standard approach in digital forensics to identify areas of unusual activity deviating from the average patter. In our case, it provides a snapshot of file lifecycles and could potentially be indicative of preparation, incubation and verification stages related to the artist's creative process (cf. Section 2.2).

Whereas each stage of the content analysis branch of investigation is computationally intensive, the metadata analysis associated with timeline generation is much quicker. It is partly for this reason that, in digital forensics more broadly and in this paper, the analysis of metadata has been employed as a scalable means of event reconstruction and extracting file relationships (e.g. [58], [87]). Given the scope of the data present on the drive and the fact that the contents are completely unknown, being able to apply data analysis techniques to the metadata and so discern the rough overall shape of the data is of great benefit for guiding the content analysis.

Date	Size (bytes)	Type	Mode	File Name
2011-07-15T12:11:05Z	1.82E+08	...b	r/rrw-r--r--	AGE_RUSHES FOLDER/AGE_ITdancebelly/dance20.mov

Table 2 - Example columns from Sleuthkit timeline: date and time of activity, file size, type of activity (…b denotes that this is a record of when the file was created or "born"), mode/permissions associated with the file, and file name.

3.3 Content analysis

Our understanding of Dwoskin's archive can be enhanced if timelines built on disk drive usage and type of activity are correlated with information about the content. Content can range anywhere from title and descriptions such as file names, file extensions, and folder names, to data units such as textual and/or video content of the digital materials on the hard drive.

File extensions can be indicative of files formats, and file and folder names can be indicative of conceptual organisation of material. Combined with temporal metadata, these can suggest an overview of different types of materials (e.g. audio/visual, text), technological choices (e.g. software) and engagement with clusters of material associated with changing themes over time. File extensions can be cross-referenced with format registries (e.g., PRONOM[15]), format validation (e.g., DROID[16], JHOVE[17]), and metadata extraction tools, toward full format identification (cf. [30]). Identifying formats not only sheds light on Dwoskin's technological environment and possible ways of accessing the information, but also locates specific types of content (e.g. audio/visual content, documents, emails) that might be targeted for analysis and research.

Likewise, the folder structure indicates topic/themes that bind files together, for example, Dwoskin's project names. Analysed in parallel along with an understanding of materials that were accessed at the same time, this could define the scope of search area for deeper content analysis for collaborations and editing cycles. In particular, the video content that can be found on the hard disks can be seen as valuable resource to better understand Dwoskin's work: changing styles, main themes, and discarded edits.

15 https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

16 http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

17 http://jhove.openpreservation.org/

The primary advantage of conducting initial analysis of the type described throughout this methodology, is that it allows for us to establish the overall shape of the data present on the drives, both in distribution and in size. This in turn is indicative, if not conclusive as to the type of content. For a large collection, it is important for the archive to establish an efficient method of sorting and prioritising the process of accessing the content and this type of analysis is an essential first step that requires few resources and little time. However, this paper aims to take the data analysis beyond the initial requirements for the archive both to provide a more detailed report for access and appraisal purposes and to demonstrate the potential of such analysis to complement the archival record and tailor it towards the potential research needs of the user.

4. Exploring Dwoskin's creative process and environment

The discussion in this section, directly addresses the research questions raised in Section 2.5. The first research question will be mostly discussed in Section 4.1, the second in Section 4.2, and third in Section 4.3. However, the interrelated nature of the research questions means that some aspects of each of the question will emerge throughout.

4.1. Patterns of personal and professional history

4.1.1. Professional filmography

Figure 2 represents the count of file activities extracted from the images (total recorded: 67,541) associated with each year, compared to Figure 3, which demonstrates the overall size of the corresponding files. Each bar on the graph is subdivided to demonstrate what proportion of the activity is present on each drive. Before delving into the analysis, there are a couple of points to note. For example, some of the x-axis entries may appear to be empty, but this is a matter of perspective – if a date appears on the graph then there is at least one

timestamp present for that date. Equally, some dates present on these two graphs will be omitted from further analysis: for example, dates displayed as 0000-00-00T00:00:00Z (0 on the graph) are not valid dates – Sleuthkit, as a policy, assigns these to activities where it could not identify relevant dates. Similarly, there are the properly formatted if unlikely dates of 1970[18] and 2036. These dates will also be omitted from the analysis in this paper.

The earliest date other than 1970 recorded in the data is 2000 on Disks 03 and 06 and the latest realistic date is 2013 present on all drives except 04. The assumption is that older disks would be first used at an earlier date and that their related activities would cover a wider range of dates thereafter. Using these assumptions, we might posit an order to the disks from oldest to newest content: Disks 06, 03, 01, 04, 10, and 02. As it happens, the order is consistent with the fact that traces of one of his last films "Age is ..." (2012) can be found as a project on Disk 02 with dates from 2011, while traces of his film "Another Time" (2002) can be found on Disk 06 and on no other drives.

[18] These timestamps seem to be tied to operating system files and so could be representative of the start of Unix time. However, to determine this for certain would take additional analysis that has not been conducted for this paper.

In terms of usage, the majority of the drives seem to be associated with focused periods of activity at specific points, possibly for specific projects. Disk 02, for instance, has almost no activity until 2011 but, at that is the dominant drive both in terms of activity count and file size. Disk 04, similarly, only seems to have been utilised to any great extent in 2005 and 2009. Despite this, in comparing the activity count of this drive, as seen in Figure 3, to the file size of Figure 4, the average file size is quite comparatively small, particularly for the years outside of indication that the activity in the later years was only on specific smaller files. Disk 10 is almost the converse of this, in that it demonstrates a small activity count but a relatively large size of file particularly within 2007. The overlay in the figures demonstrate that the dates correlate with the production of Dwoskin's film "Oblivion" (2005), three films in 2007, "The Sun and the Moon", "Phone Strip" and "Phone Portrait", two in 2008, "Ascolta!" and "Mom" and the film "Age Is …" (2012). Whilst not, at this point, conclusive, this patterning does further enforce the hypothesis that each drive, disk 01 excluded, is focused upon a specific film/set of films with these intense periods of activity indicated the period of composition, editing and/or archiving of specific projects.

In contrast to these focused drives, in examining Figure 2 it is evident that Disk 01 defies the expected pattern. In part, this is unsurprising given that it is the most ubiquitous of the drives with approximately 55% of the total activity. However, this volume of activity is not the sole explanation. There is evidence of activity in almost every year in the given period although much less so in 2006-7 and 2010 and not at all in 2000, 2009 and 2012. The

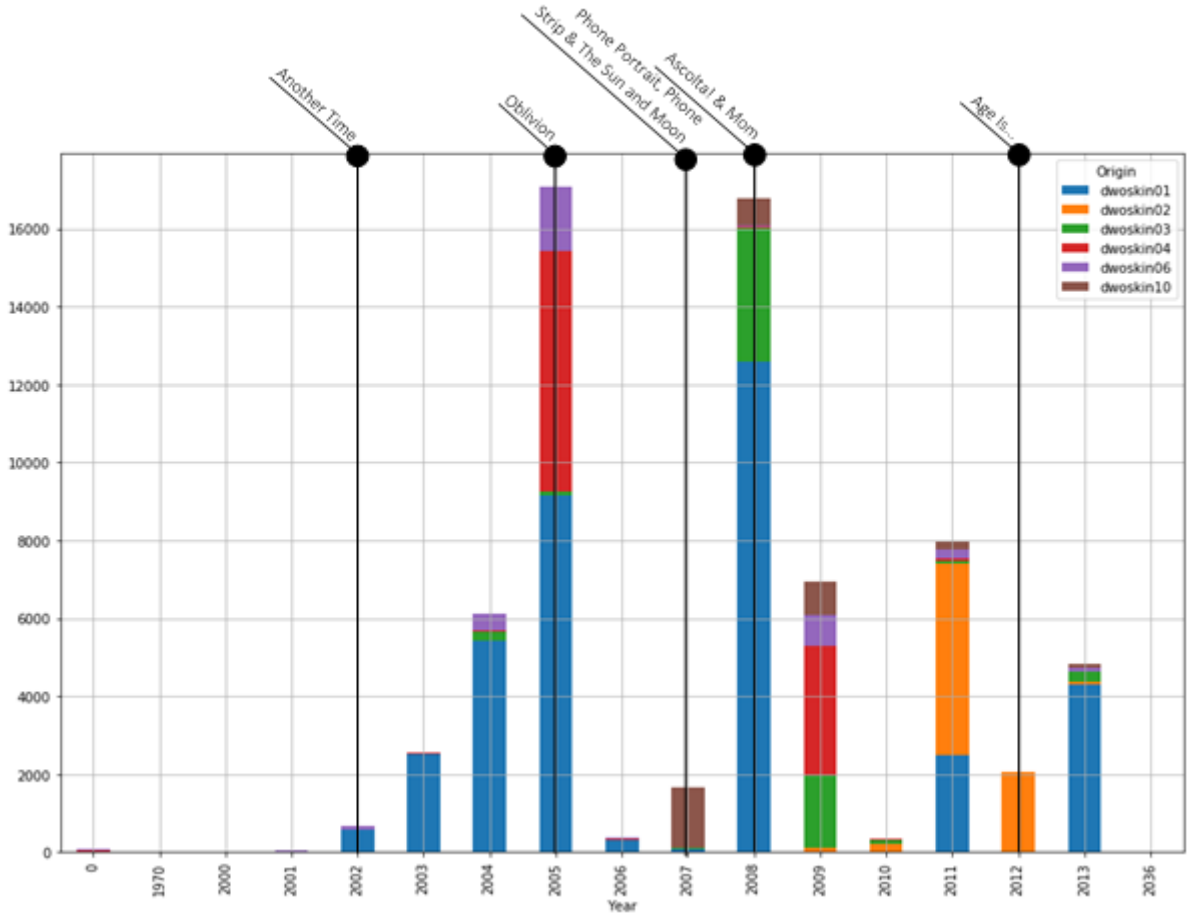

Figure 2 - Total file activity count for the six drives being analysed in this paper. The subsections of each bar represent the different drives. The overlay indicates the years in which certain films were released. Colours indicate different drives and overlay indicated film release dates – print with colour.

high level of usage of this drive, contrasts with the comparatively small file size as demonstrated by Figure 3. This observation, along with the persistent usage of the drive over time distinguishes it from other drives and suggests that it might have been a system back-up drive rather than an audio-visual content storage device.

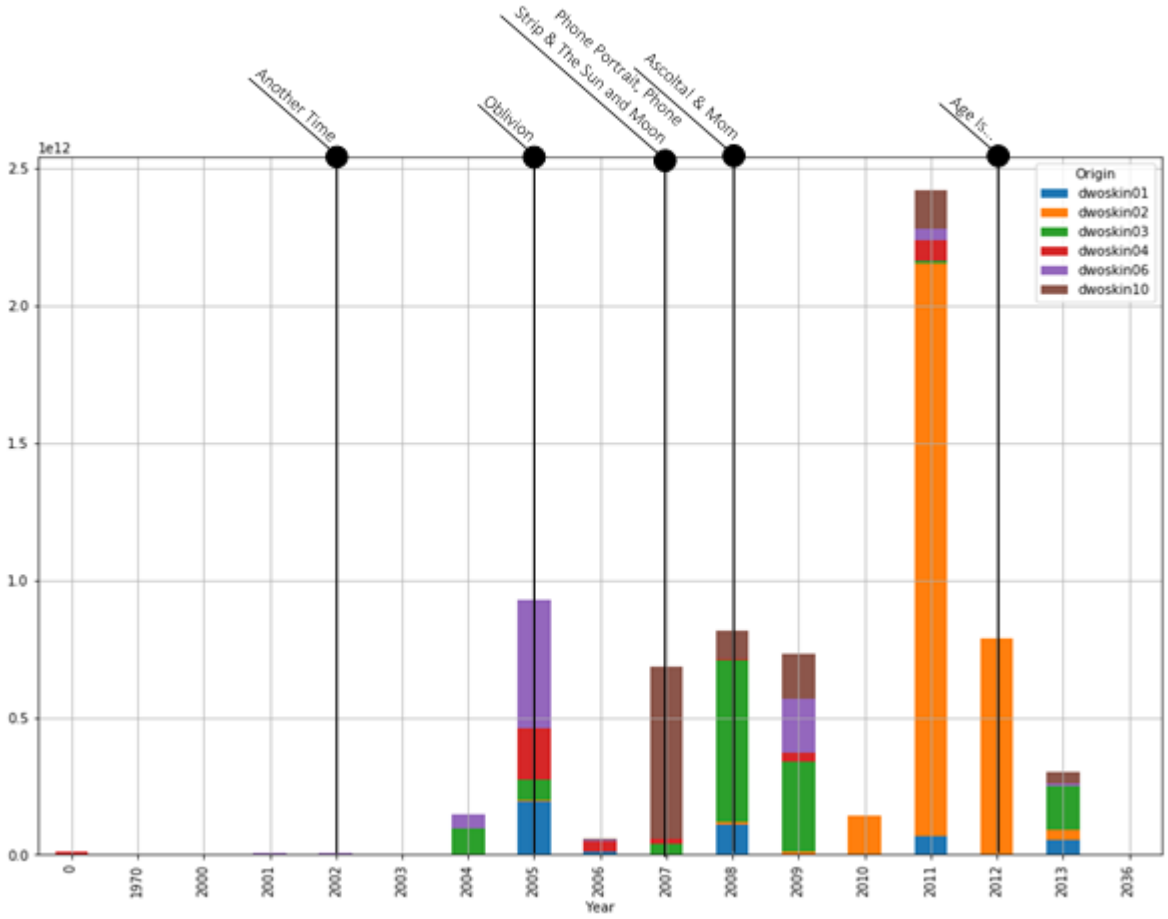

Figure 3 - Total file activity size for the six drives being analysed in this paper. The subsections of each bar represent the different drives. The overlay indicates the years in which certain films were released. Colours indicate different drives – print with colour

4.1.2 Working patterns

Given that there are potentially two different types of drive usage amongst our sample, the suspected back up/core computer drive Disk 01 and one of the content drives Disk 04 were selected for the initial examination. With the data for each set isolated from the main body (Figure 4 & Figure 5), it is clear that Disk 01 has consistent usage but with definitive peaks and Disk 04 has only two peaks, one in 2005 and one in 2009 but with other low-level activity throughout the 2003-2011 range.

What is additionally evident from the monthly breakdowns of each year provided by these more granular graphs, is that much of the activity, on both drives, is clumped into a few months of the year. Both of them have some degree of activity in every month throughout the range of the drive but, as an extreme example, there were 12,000 instances of activity on Disk 01 that all occurred in the month of September, a spike which, given the volume, can be speculated to be a backup or installation activity. Conversely, Disk 04, although with far fewer

instances of activity, does appear to follow a similar grouping in that the majority of activity in each year occurs in one month. Furthermore, there would seem to be a degree of seasonality to the working patterns in that these large collections of activity, for the most part, occur in the summer months. This is most evident in Disk 04 with almost all of the activity occurring between May and September and a far smaller, if noticeable, amount in October and November. Disk 01 displays similar chunking around the summer months between 2005 and 2011 with 2004 just outside of this range. Prior to this, the patterns seem to have been a little different with the primary focus of activity being the winter months November through to January. However, as previously observed, larger file activities are associated to those after 2004, suggesting that these earlier patterns are not substantially indicative of development of audio-visual materials.

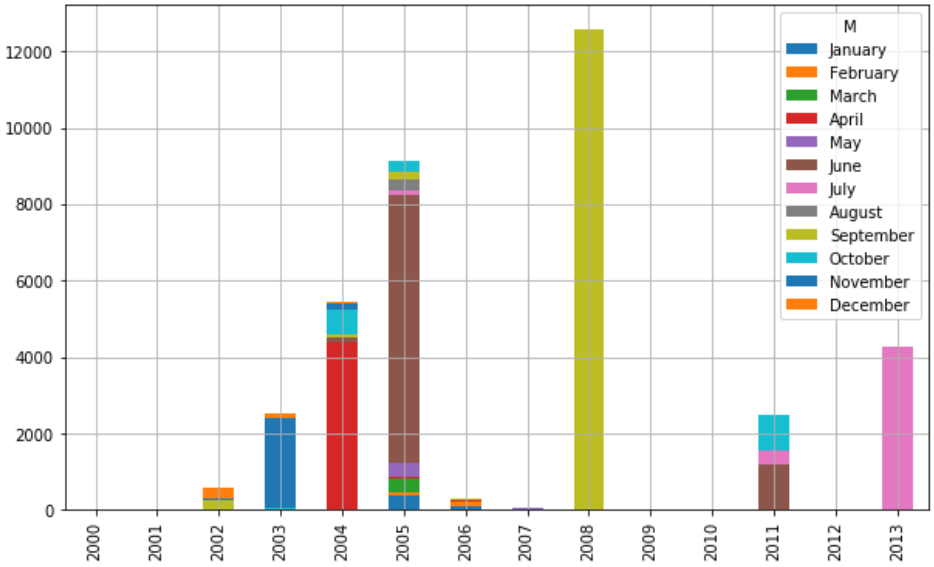

Figure 4 - Dwoskin 01 breakdown of activity by year and month. Subsections of each bar represents a different month. Months are indicated with different colours – print with colour.

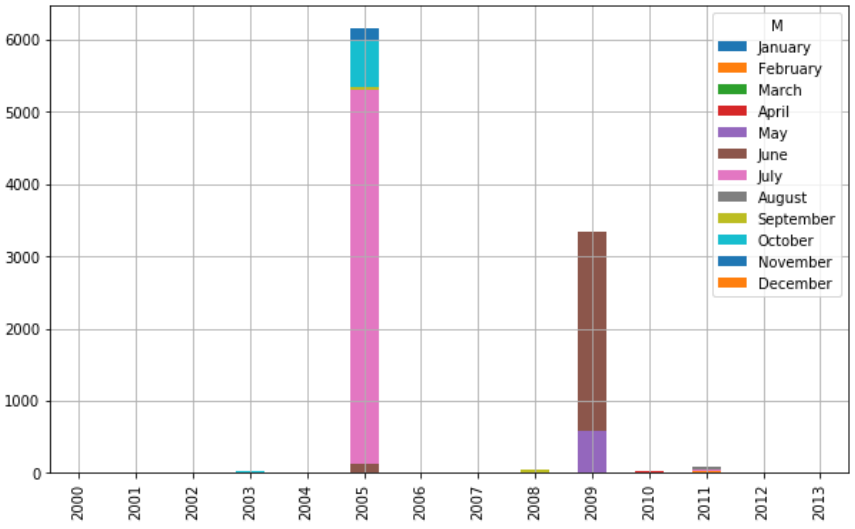

Figure 5 - Dwoskin 04 breakdown of activity by year and month. Subsections of each bar represents a different month. Months are indicated with different colours – print with colour.

In expanding the month groupings to the whole dataset (Figure 6), the seasonality demonstrated in Figure 4 and Figure 5 is shown to extend to the other drives. The large number of files present on Disk 01 does skew the structure of the data slightly, but even with this taken into account, the majority of activity is focused between May and September. February, March and December are the quieter periods across the dataset and November on all drives but Dwoskin 01. This seeming quasi seasonal workflow is more likely to be indicative of a true work pattern even in spite of the missing data from the as yet un-imaged drives, because it is consistent across the sample of storage drives. Whilst the reason for this pattern is not obvious from the metadata this is something that can be investigated by examining the types of files created and accessed during these periods as well as more traditional historical research techniques to investigate real world influences that could have affected this.

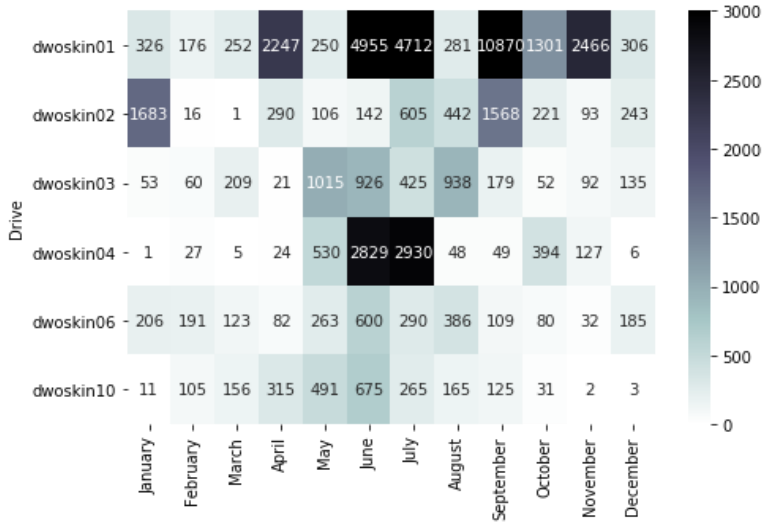

Figure 6 - Distribution of activity across months for all drives. Subsections of each bar represents different drives. Colours indicate intensity of activity – print with colour.

4.2 Stages of creative process

The graphs presented in Section 4.1 represent total file activity within a given range rather than an actual count of the files present on the drive. Another dimension can be added to the timeline analysis is the examination of the creation 'b', modification 'm' and access 'a' times (MAC times) associated with each file activity to glean general patterns of specific types of file and/or directory manipulation. Noted by Dan Farmer as both the 'most potentially valuable forensic tool in your digital detective toolkit' and as 'woefully underutilized' [33] the MAC times have numerous benefits to timeline analysis. However, they must be treated with caution. From a purely practical perspective, the MAC times are not wholly reliable both because they can degrade overtime, displaying a phenomenon referred to as "digital Alzheimer's" and because, in certain circumstances, the user can willingly change the timestamps [33]. In addition to this, they must be used in the knowledge that they cannot provide a complete picture, but rather an indication of file activity (usually only the last dates of activity for creation, modification and access are available). Whilst this measure cannot show how often a particular file was utilised, it does give a good indication as to the overall scope of a particular file's lifespan.

The Sleuthkit mactime tool that was used to extract this metadata, an example of which was presented in **Error! Reference source not found.** (Section 3) creates an entry for each MAC event date and provides the 'Type'

column to define what the date signifies. In the current dataset the type ‘b’ (rather than ‘c’) denotes the date the file was created (‘b’ stands for ‘born’). While the b timestamp is not necessarily reliable as the true time of creation, the date can serve as a relative reference point for a file, given that the change of the ‘b’ date would normally signal the change of the ‘m’ and ‘a’ dates as well. The ‘a’ time is the most changeable of the timestamps as it can be updated simply by accessing the directory of the file – part of the reason that a disk image, rather than the disk itself is used as a source for this research. The ‘m’ stamp occurs only when a file is modified and so this, compared to the b stamp, can give an indication of the active lifespan of the file, the period through which it is accessed for the purposes of updating or altering the file. This contrasts to the passive lifespan, more clearly indicated by the ‘a’ timestamp, which demonstrates interaction with but not alteration of the file. The fourth timestamp, ‘c’ in the dataset, not analysed in this paper, refers to a metadata change [33].

In separating out the different types of timestamp, it is possible to get a clearer idea of how the activity on the drives was distributed across true creation and editing activity compared to viewing and interacting with a file without altering it. Viewing first the distribution of the types of activities across the entire data set, some of the discrepancies found in the undifferentiated dataset (Figures 2 & 3) above are explained and the patterns become clearer. Figure 7 and Figure 8, for example, which show the b-time and the m-time respectively, demonstrate that for Disk 01, 04 and 06 the majority of creation and modification activity occurred prior to 2006. The other drives Disk 02, 03, and 10 were used for this type of activity predominantly between 2007 and 2012, respectively. These graphs clarify the groupings of activity mentioned in Section 4.1 and further emphasises that each drive, for the most part, had a specific period of activity so, it can be inferred, a specific body of work. In addition, the distribution of Disk 01, with a high rate of production and modification activity consistently across 2003, 2004, and 2005 further differentiates it from the other. The lack of similar activity across other years would suggest that there will be other drives within the broader collection, as yet unidentified, which were used for the same purpose as this one, or that the type of activity on this drive was discontinued.

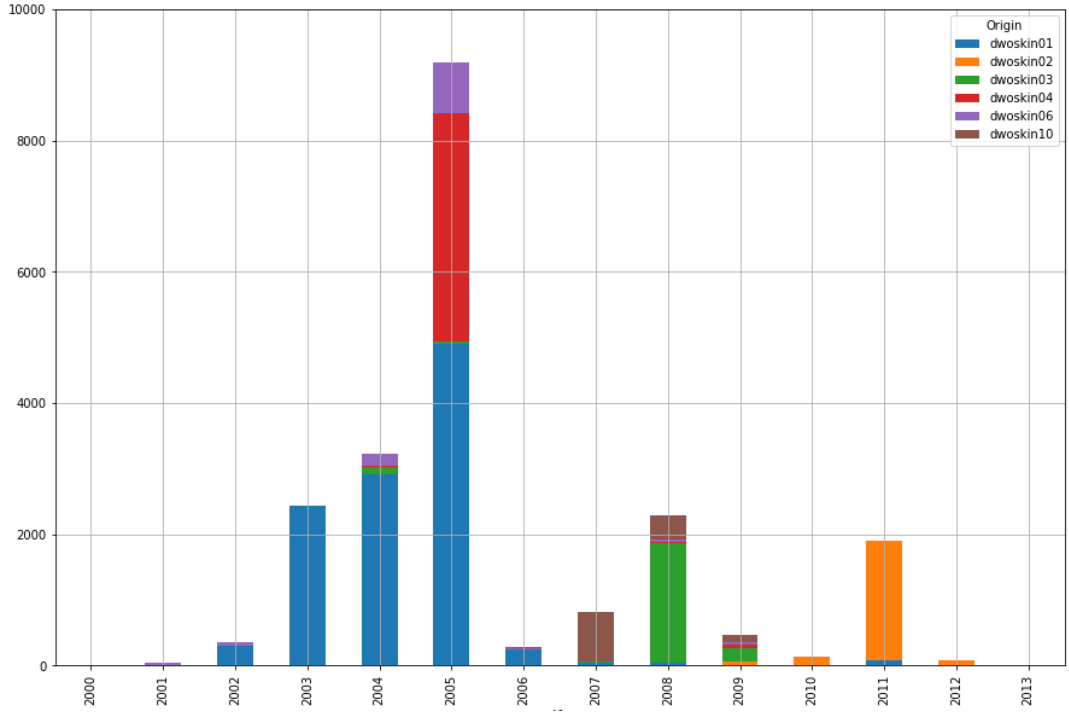

Figure 7 - Graph representing the 'b' timestamp across the entire dataset. Subsections of each bar represents a different drive. Colours indicate different drives – print with colour.

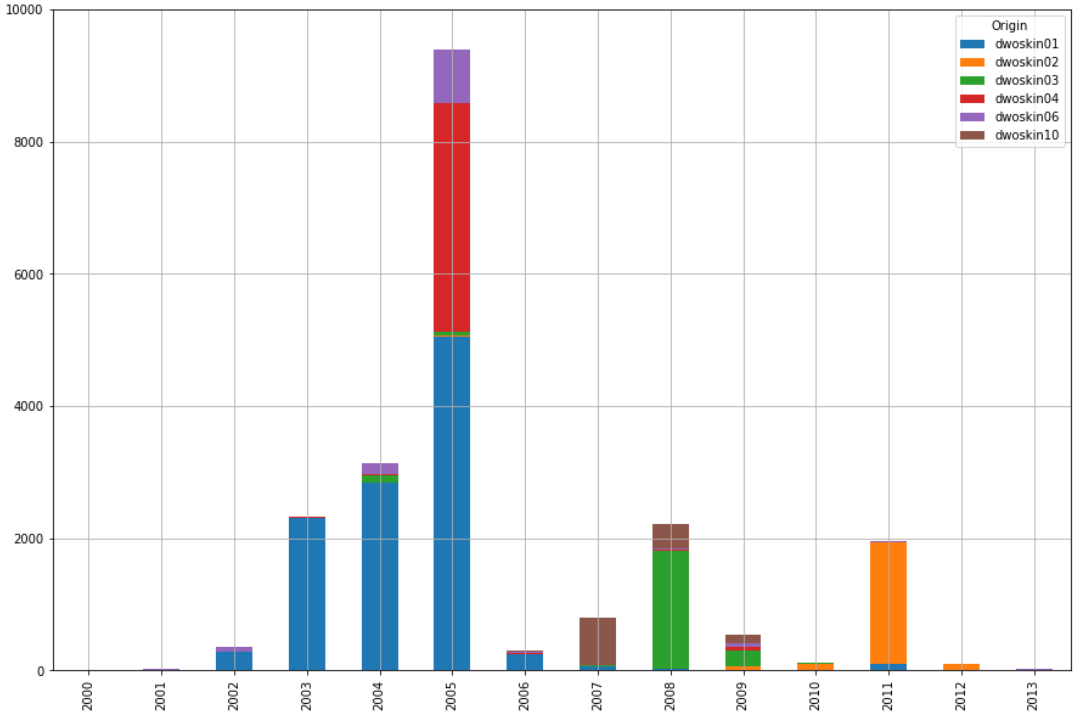

Figure 8 - Graph representing the 'm' timestamp across the entire dataset. Subsections of each bar represents a different drive. Colours indicate different drives – print with colour.

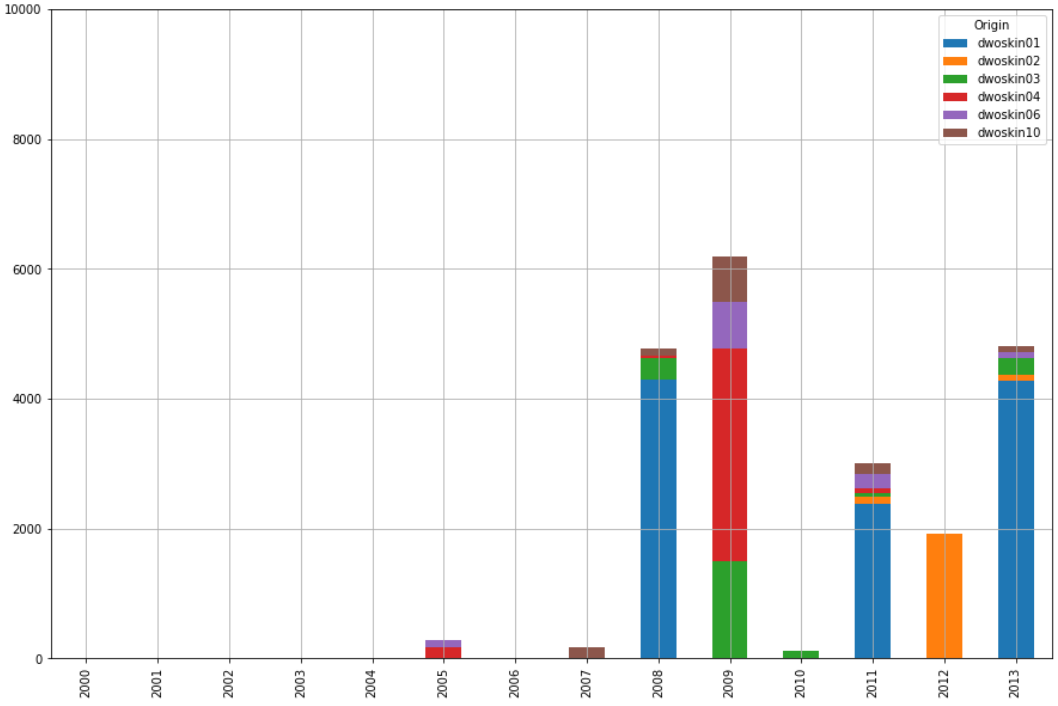

Figure 9 - Graph representing the 'a' timestamp across the entire dataset. Subsections of each bar represents a different drive. Colours indicate different drives – print with colour.

With regards to the drives with more focused usage, similar gaps can be identified by a comparable lack of creation and modification activity before 2005, in 2006 and 2009-2010. Whilst it is possible that these years were fallow years in Dwoskin's work cycle, given the consistently high rate of activity in other years, this would seem unlikely.

In comparing Figure 7 and Figure 8, it is intriguing how closely the 'b' and 'm' timestamps correlate to one another with an overall standard deviation of 99 days between the two activities. This difference would indicate that, on the whole, the process between creating the file on the drive and the final date for modification of that file is relatively succinct. It is at this point that we begin to see the individual activities present on Dwoskin's drive coalesce and begin to form a narrative of events. Taking Disk 04, for example, this drive was modified almost entirely in 2005 with only a very small secondary series of creation and modification in 2009. This focused usage would suggest that this drive, more than any other had a particular use that, when completed, was not returned to or disturbed. When this is compared to Figure 9 presenting activities associated with access, it is evident that, although some files were not accessed again after 2005, a significant number of files were last accessed within 2009 potentially as a part of a review process for Dwoskin's material, potentially as a part of preparation for another project. On a smaller timescale, this pattern can also be seen with Disk 02, in that the bulk of the creation and modification activity occurs in 2011, but there is a similar level of access of that material in 2012. Aside from a very small quantity of files that were accessed most recently in 2005 on Disk 04 and 06, all of the access dates are post 2007 with the majority in 2008-2012. In fact, compared to the 99 days standard deviation between the 'm' and the 'b' timestamps, there is 1007 days standard deviation when examining differences between a file's a-time and its m-time.

It is within this patterning of the types of timestamps that the first hints into Dwoskin's transition between the four stages of creative process can be found. Whilst not conclusive, the distinction between the born and last modified dates for a particular series of files that constitute a project could be indicative of the movement through the third and fourth stages. This period represents the steady solidification of the stable network, as seen in a coalescence of disparate image, video and audio files into the computer environment, and the process by which this material is sorted, selected and reformed into the final product. The first two stages of the process are more difficult to pin down within a computer environment as they are far more likely to have occurred in response to real world stimuli or at the point of filming where the digital and the real intersect. However, as the digital record we have for Dwoskin's work includes both personal and professional content a certain degree of the initial associations may be found in conducting concurrence analysis of content present on the drives but not pertaining to the project in question. Whilst the times stamps are indicative, these associations can only be explored fully with, for example, examination of the file names/types as discussed in section 4.3 and 4.4 below, or, as will be done in future papers, through a concurrent content and timeline analysis.

As a final note, it is interesting that, in a comparison of Figure 7, Figure 8 and Figure 9, nearly all of the activities in 2013 are associated with access. It is known that Dwoskin passed away in June 2012. The data activity is consistent with the conjecture that, although the workable date range of the drives encompass some activity in 2013, this must be attributed to system process or other users. This is supported by the observation that the few modification timestamps found in 2013 seem to be limited to system files, probably triggered when access took place.

4.3 Materials and environment

4.3.1 File type clustering

File extensions, whilst the most evident indication of file type to be found in the file metadata, can be problematic in terms of their reliability, particularly for older files. On older computers, file extensions were not automatically provided when saving a file, and, equally, they were easy to alter, should the user wish to. Nevertheless, they are easily accessible, when present, and so readily available to explore for what they can reveal about the dataset.

Figure 10 - Word cloud representing the frequency of extensions within the dataset

When exploring the frequency of the extensions present in the timeline metadata across all the drives, presented in Figure 10 as a word cloud, it is evident that the highest frequency extensions are .mov, .sd2, .jpg, .html and .gif, with other notable extensions being .aif, .wav, .l, .tif and .htm. Overall, these extensions are expected within the context of a filmmaker's computer. The .mov extension, the most popular extension is usually associated with Apple QuickTime Movie software suggesting that this was one of Dwoskin's key tools for viewing his work. The .sd2, .l, .aif, .wav extensions are audio files, the first pertaining to Sound Designer II and the latter three from more generic sources. Image files are also very popular with the .jpg, .gif, .tif all falling within this category. These high frequency extensions were expected on the drives of a filmmaker, but, unexpected, was the prominence of the web technology structured text files with .html and .htm extensions. Upon a closer inspection of these

surprising files, it would appear that they are mainly composed of support documents that accompany installed software, but it is certainly something that requires a deeper investigation at a later point.

Intriguingly, the lower frequency extensions, as well as including some common and expected entries such as .dmg, .png, .mp4, .par, .doc and .pdf, include some very unusual extensions, or familiar extensions with unusual additions. Some of these are system (e.g the Mac Spotlight files such as .indexcompactdirectory) or programme specific files (e.g Excel's .xla) but others such as the .aiff0001kgwv and .wav0001kgwv could be indicative of user or program intervention of the file name. The extension analysis has much more information to yield, however this will have to be returned to in future papers. The key thing to note from these extensions is that, although an easy and quick method for establishing the types of files present on the drives, such statistics are not necessarily reliable or comprehensive, as demonstrated by the fact that less than a third of the data set originating with the six drives have recognisable extensions. Equally, whilst the detailed insights that are possible from extension analysis are useful, the technical skill and research required to effecteively achieve this is prohibitive with larger datasets and is not ideal for gaining an overview of the working environment.

Whilst the extensions are associated with file formats and software and, as such, can be a useful gateway to understanding the potential file types present on the drives they pose a number of difficulties, particularly for legacy software, as the extensions were not always automatically assigned and could easily be altered by the user. Therefore, to complement the extension analysis we analysed the drive content using the aforementioned UK National Archives DROID tool. This tool was of particular use for its ability to identify the software first through MIME type and then by extension. These are linked to the PUID database which contains an accompanying categorisation into a general file type such as video, text, audio etc. However, as demonstrated by Table 3 the categorisations provided are more certain, without the ambiguity of file extension analysis and, equally, a larger number of the files have been identified, approximately 55%.

Classification	Total Frequency	Dwoskin_01	Dwoskin_02	Dwoskin_03	Dwoskin_04	Dwoskin_06	Dwoskin_10
WordProcessor	128	104	9	3	0	0	12
Audio	2002	892	291	119	613	69	18
Presentation	30	24	0	0	4	0	2
Unclassified	13282	7366	814	1097	2616	742	647
Page Description	37	28	3	1	1	0	4
Spreadsheet	8	8	0	0	0	0	0
Text	2537	2510	3	6	11	2	5
Aggregate	329	276	7	13	11	11	11
Image	2739	1508	336	133	92	366	304
Font	1	1	0	0	0	0	0
Video	6945	454	2090	2376	627	509	889

Table 3 - Frequencies of files within each category present on each drive

Table 3 - Frequencies of files within each category present on each drive can be utilised to gain an idea of Dwoskin's work environment. The high frequencies of file types which can be categorised as formats linked to filmmaking indicate a promising wealth of contents for future analysis. Equally, it is intriguing that there seems to a variety of other media types. For example, the text category which refers to formats such as .html and are heavily weighted towards Disk 1 are indicative of, for example, software files. This combination of file types

could reveal information about the context of Dwoskin's life, influences and practices, particularly in terms of his creative process. Within the context of the four stage model [65] that typifies creative process research, the artefacts of film itself, for example the stills, the audio/video cuts and the final version which combines these disparate images, will most probably inform primarily on the latter stages of the model, the formation of a network, the stabilisation of that knowledge and then the sudden illumination that follows. The alternative media forms, the word processer documents, potentially some stills, could serve to enlighten upon the earlier stages, suggesting those initial sparks of inspiration and the period of gathering material.

The modified 'm' and access 'a' timestamps for these files have been displayed on Figure 11 and Figure 12 with the intention of establishing a pattern for Dwoskin's working practices, with a focus on the types of files he used at a given point within a project and throughout his career. The groupings of particular interest are the video orientated file types can reveal the order in which each aspect of the project was explored, which format was worked on first, whether there is a temporal pattern to this working process and whether this process was uniform or varied. The first stage, conversely, the period of rumination and discovery, could possibly be seen in the earliest edits or images associated with a concept, but just as readily it could arise from the contextual happenings of Dwoskin's life: an email conversation with a friend, for example, or the love of a particular piece of music. It is in the incidents of concurrence between these more mundane items and the files that pertain his work that, perhaps, the foundation stones of Dwoskin's creativity can be found. Whilst the details of this are the topic for future, the patterns of usage are analysed here for any patterns of concurrence between the types of material on the drives. In addition, when compared to the MACb timestamp graphs, in Figure 7, Figure 8, Figure 9, this waxing and waning of file types could perhaps be indicative of Dwoskin's growing confidence with digital video files, given the steady growth in their number throughout these years, or equally, to a shift in software which demanded a higher number of video or image files in the later years but diminished the number of audio files required for a project.

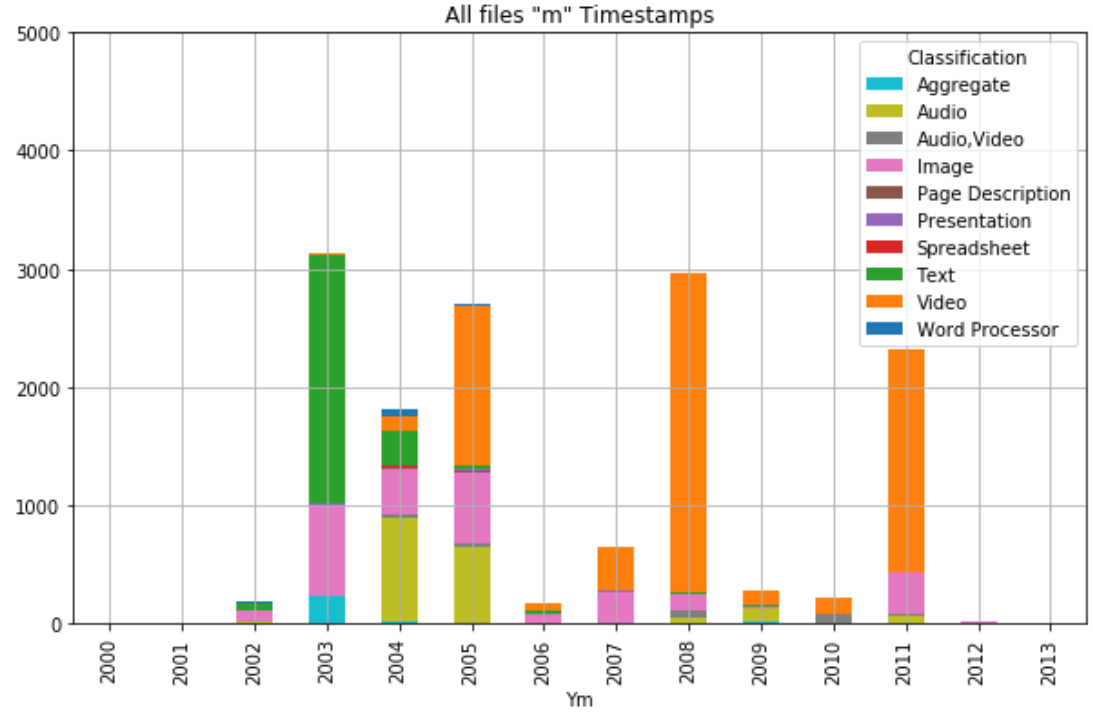

Figure 11 - Graph representing the 'm' timestamps for the file classifications identified by DROID. Colour indicates file type classifications – print in colour

If this is the case, then the file type distribution in Table 3 - Frequencies of files within each category present on each drive could be used as a further indicator of each drive's age, or, at least, order of usage.

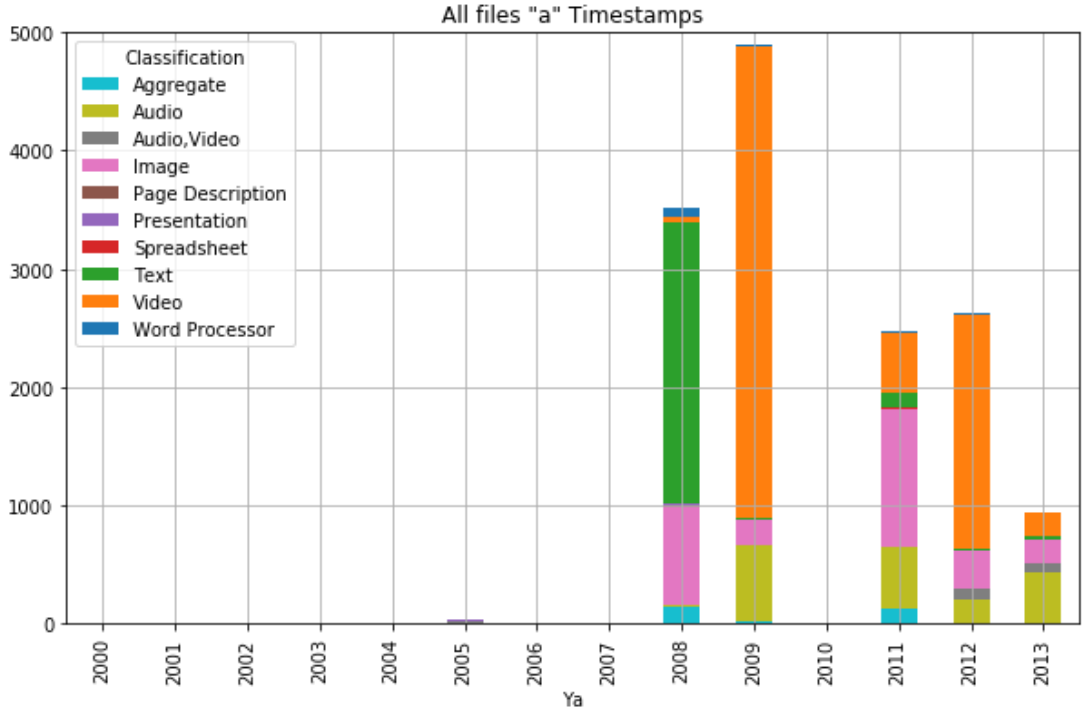

Figure 12 - Graph representing the 'a' timestamps for the file classifications identified by DROID. Colour indicates file type classifications – print in colour

A closer examination of Figure 11 & Figure 12 demonstrate the distribution of modified 'm' and accessed 'a' timestamps with respect to the file types present within the drives. As discussed in section 4.2 above, the distance between the types of dates indicates the lifespan of a particular file, in particular the distance between the born, the modified and the accessed timestamps give an indication firstly of how long a particular file was worked upon, and then at what point, if at all, it was returned to. , which displays the m stamps supports the hypothesis that Dwoskin work shifted in weight from primarily image and audio based content towards more video centred content. It is particularly interesting that after 2006 there is little to no audio based files that are being modified, but each file with an audio type is accessed most recently between 2009 and 2013. This is in fact true for all of the categorised files, as demonstrated by Figure 12. Bar a negligible quantity of files whose last access date was 2007 or before, all of the files have been accessed post 2008, mirroring access pattern from figure 13 above, but more extreme. Interestingly, the 2013 bar of the access graph for the classified files is very different to that representing the whole collection, with only about a quarter of the accessed files classified into audio, image or video files. This difference may suggest that much of the access in 2013 was incidental to the recovery of film material.

4.3.2 File path exploration

An additional method of exploring Dwoskin's working environment can be conducted through an examination of the file paths. Exploring the file and the folder names within the drives, it is evident that there is a mixture of software produced content, such as the folders names 'Audio Render Files', 'Capture Scratch' and 'Autosave Vault' which form part of the Final Cut Pro production process, a tool that Dwoskin is known to have used. Alongside these, there are the folders and files which would seem to have been named by Dwoskin himself. For the purposes of event analysis, the file names can be utilised for understanding the type of actions that may have been related to events, for example, triggered by projects and/or software.

From the folder and file name analysis of Disk 04, we see that there are some familiar names, such as 'OBLIVION' which are referenced in the filmography seen in Figures 2 and 3. Conversely, we have also found a number of potential project names which are not explicitly mentioned in the filmography, such as 'Titanic', 'fade' and 'Ghetto Project 1'. According to the timeline, the files related to 'fade' appear to be clustered with those pertaining to 'OBLIVION' suggesting a potential relationship in the creative development of these projects. This opens up the possibility of exploring additional relationships between these kinds of concurrent activities and their relationship to Dwoskin's creative process.

Even without content analysis, file and folder names, especially when it is combined with what we know about file types (Section 4.3) have much potential to provide insight into technology at the heart of project events (e.g. audio and/or video; Final Cut Pro and/or Logic Pro), when these were used for different stages of creative practice.

5. Discussion

The above findings, whilst only the tip of the iceberg in terms of the potential that can be drawn from Dwoskin's digital archive, have revealed numerous areas for discussion with regards to archival practice for personal digital archives of artists generally and film makers specifically, for research into creative practice and, for practitioners who are concerned with the preservation of their work for archival purposes. The key issue for a digital archive, particularly one of this size, that this research is intended to address is that of access to the data. In this scenario, the term "access" is complex and multifaceted, covering a variety of activities from both an archival and a user perspective. The first of these, which is the issue of legal access (e.g. Data Protection, Intellectual Property Rights etc.) is not discussed in this paper as, at present, the archive is closed for review.

The second type of access, foundational to the creation, maintenance and usage of an archive, concerns the identification of the content. The research presented within this paper highlights how digital forensics and data analysis can be employed to broadly identify the primary purposes of Dwoskin's hard drives, whether that be project based or backup, and equally opens up the route to identifying the specific projects within the dataset, through file type and file path analysis. The intense periods of activity have been mapped to known film release dates, further narrowing down disk usage and content identification, and, finally, this research has revealed the possibilities for tracking the different stages of Dwoskin's creative process through the MACb timeline analysis, further aiding in identifying key areas of interest within a sizeable dataset. This methodology therefore represents the beginning stages of developing sorting, searching and visualisation techniques that will aid both in identifying and classifying material, and in opening up a clear, multi levelled, overview of the archive to the end user.

The third type of access refers to the practical ability to view and, potentially, interact with the content within the archive. Through our research, we have begun to build a profile of Dwoskin's working environment, in particular the types of files he prioritised at different points within his career. As a subset of this analysis, we have also been able to identify, some of the specialist, proprietary and/or legacy software that will be needed to access and work with the content on the drives. This allows us to gain an understanding of the technical specifications required to access the content whether that be by building a legacy virtual machine or acquiring legacy software. As with the second type of access, where this research offers the greatest advantage, however, is by allowing this analysis to occur across the whole dataset with minimal usage of resources whilst also producing additional data that can support for both the archivist in the creation and management of the archive and the user in conducting research.

In the coming section, each of these implications will be explored in greater depth.

5.1 Implications for the archive

Even in analogue archives challenges arise when there is a lack of prior knowledge and understanding and when the acquisition process affords no opportunity to make an adequate assessment: a broader problem characterised by Meissner and Greene as "the problem of too much stuff in repositories"([42]). This is particularly the case when original order has been lost, as is the case with Dwoskin's own analogue archive. In the case of hybrid archives, a lack of order in the analogue archive compounds difficulties in understanding the digital and the workflow on this project involves constant reference between these two information sources.

The processes and practicalities involved with acquiring and archiving personal digital archives are still very much in their infancy. Many institutions have a digital archiving workflow, but these are often adapted from analogue procedures and thus far there has been little scholarly discussion into the describing and cataloguing of such material ([25][60]). Nevertheless, it is well established that the ideal digital archive acquisition process involves discussion with the donor and so a comprehension of the contents and usage of the material ([60][84]). However, there are numerous cases where this is not possible. This may be because the donation was made post mortem or it was acquired from another third-party source or potentially because the size or age of the archive means that the donor themselves are unaware of the complete contents (eg.[7][27][10]). The situation arising from this is an archival collection of unknown size, unknown content and unknown value which could take an unrealistic number of man hours to sort manually, let alone to make the collection a viable source for future research. What this paper has demonstrated is that, through a relatively simple series of forensic analyses and metadata manipulation, it is possible to establish not only the scope of the data in terms of the size and number of the items present, but also to gain a comprehension of the landscape of the contents, by which we mean an idea of any overarching patterns, or large features, as well as a comprehension of the finer details. In particular, this landscape produces an idea of any concurrences within the data. For example, the temporal concurrences of materials which are otherwise distinct could indicate a synonymy of thought or an unbeknownst association between a set of files. Equally, in terms of organisational structure, the manner in which Dwoskin structured his work, the positional or relational concurrences can reveal the connections that Dwoskin himself maintained between the files, and the particular cues and memory paths he utilised to find items within his work. The third

type of concurrence is technical concurrence which consists of the file types, their locations and their associated tools.

Alongside providing opportunities to contribute to existing research into digital preservation strategies for legacy material, such information lends itself not only to the process of auditing the contents of the archive for cataloguing and sensitivity reviews, but also, when utilised in conjunction with historical records, can help to ensure the fidelity of the archival record. This is a particular concern for digital metadata which can be highly volatile (as discussed in section 3.2) and also difficult to ascribe to a specific user [46]. A keen example of the importance of this can be seen in the presence of the 2013 timestamps on the drives. Dwoskin, it is openly known, passed away in 2012 and, as such, could not possibly have created these records in the archive. It is possible, however, that, even without prior historical knowledge to accompany the archive, the data that profiles Dwoskin's working pattern, or indeed that or any computer user, can be utilised to highlight abnormal data. For instance, the 2013 data occurs within a very localised period of time, very different to Dwoskin's usual periodic behaviour and so should be flagged for further investigation. This case presents a relatively simple case of user segmentation, contrasted to cases of multiple user computers. However, it is essential to be aware and take account of such potentially misleading records as they mask the true archival record, overwriting the true timestamps without hope for retrieval. Nevertheless, beyond this additional safeguarding of a particular record's authenticity, or lack thereof (a consistent concern with digital archives [46]), this user differentiation technique can add an intriguing dimension to the archive. Particularly in instances where the computer was accessed postmortem, these anomalous records demonstrate which files were of significant enough interest to a family member/friend to search through the drives for access to it.

Another future area for discussion arising from this research is the potential of forensic metadata analysis for research into and preservation of legacy software. On a technical level this research contributes to knowledge of how particular programmes or types of usage can be identified from the metadata and from the patterning of activity, therefore providing practical and otherwise hard to find support for the preservation and use of legacy software. For example, additional examination into the file specific metadata may reveal details about a particular file type or extension that has yet to be recognised within the comprehensive but still growing PRONOM project and associated DROID software. The need for research such as this can be interpreted from the DROID analysis that was discussed above in section 4.3, and the fact that only about a half of the files contained on the computer were recognisable from the PRONOM database. The period in question represents such a rapid and extensive evolution in technology, one that is still continuing to this day, that it may take some time to ensure a complete record of the legacy software from that time, but it is hoped that the Dwoskin project can contribute to this in some way.

5.2 Implications for creative research

The Dwoskin digital archive, as mentioned in the introduction to this paper, represents a unique opportunity for those involved in creative research, but also for those who wish to utilise digital archives more generally. As mentioned in Section 2.1, the designated community for this project consists of those whose research interests are in creative process and how this intersects with moving images and sound and/or disability and the body. Such a community presents a diverse range of research possibilities, across a variety of media and an equally diverse

associated skillset. The potential user groups for this archive include, but are not limited to, the technical, the historical, the theoretical and the artistic, each of whom would be concerned with a different degree of access to specific items of content rather than the collection at large. Of particular interest to the designated community, would be the points of intersection that occur between the disciplines and the potential for interdisciplinary research in light of this. The analysis conducted in this paper provides the first stages of a fully searchable and accessible archive intended to provide for each of these areas.

For such a large collection, which is otherwise inaccessible to the user as much as to the archivist, the inclusion of the forensically acquired timestamp and filename/type metadata with the archival record encourages access from a variety of potential researchers and minimises the need for exploratory search behaviour ([51][99]) by making the contents of the drive as open and accessible as possible. The inclusion of the metadata fields aids in increasing the accessibility of the archive, allowing a potential user to find data that is only of a specific file type, containing a specific string of words or, within a certain date range. This scope for accessibility is enormous allowing for examination of the archive in any degree of detail, whether it be the minutia contained within a single frame of a film or the establishment of an overarching pattern of behaviour spanning more than a decade of work. Each level of access provides a different perspective on Dwoskin's creative process and the research that accompanies it, the first stage of which has been presented in this paper, aids in the establishment of a model for analysing this process. Such information is essential for the end-user, no matter what their research may be, to understand the context surrounding Dwoskin's digital legacy and the particular profile of how he used his digital tools, how he approached a particular concept and how he managed his work. Of particular use, is the potential of this archive to reveal Dwoskin's creative process from the inception of an idea to the finished piece. The digital archive represents a unique interchange between the physical and the digital world. Dwoskin's influences and real world experiences, that ineffable inspiration that drove him to create his work, are captured during the filming process and then are transformed using the computer to create a film, an entity that is so often considered as a single entity but is actually the summation of the creators' vision, skill and taste. By being able to view the editing process, the period between real world images/sounds and the finished product, a great amount can be revealed about how Dwoskin engaged with his media, his creative choices and potentially, the elucidate on the context that surrounded those choices.

6. Conclusions and future work

In this article we have established digital forensics as a methodology for researching creative practices and for enhancing archival practices. In particular, we have shown the value of timeline analysis (discussed in Sections 2 and 3) in determining the working patterns of a film maker's creative practice and, as such, augmenting the finding aids available for navigating the archival content.

This methodology of combined forensic analysis and (meta)data extraction has demonstrable implications for informing the manner in which digital archives can be acquired, explored and presented for access with a particular focus on the designated community (see Section 5). We have demonstrated, in this paper, how timelines can be constructed from file activity timestamps (Section 3) and associated to actions of a film maker (Section 4). These timelines support analysis into how Dwoskin used different drives for different purposes (storage, projects, back-up), and when he worked, for example, mapping his productive periods, and through doing so establish working

patterns in relation to modification, access and creation of files. Such exploration improves overall knowledge of the collection and aids close analysis of activities related to his creative practice and working environment as well as personal/professional history.

The timeline analysis presented in this paper introduces the possibility of event detection and retrieval (cf. [94]) as a way of researchers and practitioners engaging with special collections and archives in the digital. In this paper we demonstrated how periods of activity detected in a forensic analysis can be mapped to known project events (e.g. see Section 4.1.1). It could also be possible to use patterns of activity to detect unknown projects or to consolidate conflicting information. We can also ask questions about relationships between project events found to be concurrent (e.g. see Section 4.3.2) and influences of the working environment on these project events.

More generally, the correlations between different file activities can be formulated in the context of creative process. The scope of this type of analysis is valuable both for the designated community and beyond. Beyond the unique insights it can provide into an artist's working practices, it can also contribute to our understanding of, for example, how digital tools were used throughout the given period and equally, how these tools can be analysed, preserved and used in a modern context. For example, as mentioned in Section 3.3, the video footage can be seen as a rich resource for analysis since it allows us to learn more about the work of Dwoskin. Building on prior work conducted in the area of multimedia analysis, and combining this with timeline analysis, we aim to investigate further what the footage reveals about his filming and editing styles over the years. As common in the field, we will focus our investigation on a shot-by-shot level. The first challenge includes identifying shot boundaries between various scenes. Shot boundary detection has been studied for a number of years, and we intend to employ effective methods, e.g., as outlined in [68]. The first research question that we want to study here is whether Dwoskin's work consists of a consistent number of shots per film. Another challenge is to identify keyframes, i.e., representative frames for each shot. An analysis of these keyframes can reveal more about the style of Dwoskin's work. For example, a low-level content analysis could be performed to distinguish between indoor/outdoor scenes on a very basic level, or even higher-level concepts such as physical objects that were recorded. We hypothesise that such automated analysis can help us to track changes in his filming style over time. In particular we intend to exploit recent advances in concept classifiers employing state-of-the-art deep learning methods. For example, the CAFFE concept detector [56] can allow us to generate suitable content labels with potentially high accuracy.

The timeline analysis, whilst opening up the archive in a previously unexplored manner and providing a firm foundation for other research, does little to explore the 'why' behind an artist's work. Why did the artist create, modify or access the files in this way? Why was that file and not another of interest? It is only possible to explore these 'why' questions, the ones which examine the fine details of Dwoskin's creative practice, by exploring the content of the drives and, for greater insight, combining this with the contextual research which forms the other branches of this research project. The metadata analysis presented within this paper has provided an essential infrastructure for Dwoskin's digital legacy and it is because of this that we can now begin to explore and enrich the story told by the archive. As a promising research direction, we argue for the exploration of Bayesian reasoning or related probabilistic inference frameworks to identify dependencies and relationships between different types of content, timelines, and conditions.

Acknowledgements

This work was supported by the Arts and Humanities Research Council [AH/R007012/1].

References

[1] Alfeld, M. and de Viguerie, L. Recent developments in spectroscopic imaging techniques for historical paintings - A review, Spectrochimica Acta Part B: Atomic Spectroscopy, Volume 136, (2017), pp.81-105, ISSN 0584-8547,DOI: 10.1016/j.sab.2017.08.003.

[2] Alípio M. Jorge, Ricardo Campos, Adam Jatowt, Sérgio Nunes, Information Processing & Management Journal Special Issue on Narrative Extraction from Texts (Text2Story): Preface, Information Processing & Management, Volume 56, Issue 5, (2019), pp.1771-1774, ISSN 0306-4573, DOI:10.1016/j.ipm.2019.05.004.

[3] Alonso, O. Tremblay, S.-E. and Diaz, F. Automatic Generation of Event Timelines from Social Data. In Proceedings of the 2017 ACM on Web Science Conference (WebSci '17). ACM, New York, NY, USA. (2017) pp.207-211. DOI: 10.1145/3091478.3091519

[4] Althoff, T., Dong, X. L., Murphy, K., Alai, S., Dang, V. and Zhang, W. TimeMachine: Timeline Generation for Knowledge-Base Entities. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM, New York, NY, USA. (2015) pp.19-28, DOI: 10.1145/2783258.2783325

[5] Amabile, T. M. Creativity in context. Boulder, CO: Westview (1996)

[6] Ansah, J., Liu, L., Kang, W., Kwashie, S., Li, J. and Li, J. A Graph is Worth a Thousand Words: Telling Event Stories using Timeline Summarization Graphs. In The World Wide Web Conference (WWW '19), Ling Liu and Ryen White (Eds.). ACM, New York, NY, USA (2019) pp.2565-2571. DOI: 10.1145/3308558.3313396

[7] Bakker, T., "Floppy disks and FireWire devices: Towards an understanding of the shifting nature of musical sketch material" Swedinsh Journal of Music Research Vol. 99 (2017), pp.21-35

[8] Barker, R. Trapped in the Filter Bubble? Exploring the Influence of Google Search on the Creative Process. Journal of Interactive Advertising 18:2 (2018), pp.85-95.Barmpatsalou, K., Cruz, T., Monteiro, E. and Simoes, P.. Current and Future Trends in Mobile Device Forensics: A Survey. *ACM Comput. Surv.* 51, 3, Article 46 (2018), pp.31. DOI: 10.1145/3177847

[9] Barros, C., Lloret, E., Saquete, E. and Navarro-Colorado, B. NATSUM: Narrative abstractive summarization through cross-document timeline generation, Information Processing & Management, Volume 56, Issue 5, (2019), pp.1775-1793, ISSN 0306-4573, DOI:10.1016/j.ipm.2019.02.010.

[10] Becker & Nogues, 'Saving-over and Over-Saving and the future mess of writer's digital archives: a survey report on the personal digital archiving practices of emerging writers', The American Archivist, Volume 75 (2012) pp.482-513

[11]Belshaw, Scott H. (2019) "Next Generation of Evidence Collecting: The Need for Digital Forensics in Criminal Justice Education," *Journal of Cybersecurity Education, Research and Practice*: Vol. 2019 : No. 1 , Article 3.

[12] Bettivia, R. S., The power of imaginary users: Designated communities in the OAIS reference model. Proc. Assoc. Info. Sci. Tech., 53: (2016), pp.1-9. DOI:10.1002/pra2.2016.14505301038

[13] Botella, M., Nelson, J. and Zenasni, F. It Is Time to Observe the Creative Process: How to Use a Creative Process Report Diary (CRD). The Journal of Creative Behavior 53:2 (2019) pp.211-221.

[14] Boyd, C. and Forster, P. Time and date issues in forensic computing—a case study, Digital Investigation, Volume 1, Issue 1 (2014) pp.18-23, ISSN 1742-2876, DOI:10.1016/j.diin.2004.01.002.

[15] Brett, J. and Jones, J., "Persuasion, promotion, perception: untangling archivists' understanding of advocacy and outreach", Provenance: Journal of the Society of Georgia Archivists , Vol. 31 No. 1, (2013) pp.51-74

[16] Brylla, C. The benefits of content analysis for filmmakers, Studies in Australasian Cinema, 12:2-3 (2018) pp.150-161. DOI: 10.1080/17503175.2018.1540097

[17] Cagle, M. A general abstract–concrete model of creative thinking. Journal of Creative Behavior, 19, (1985) pp.104–109.

[18] Cao, Joachims, Wang, Gaussier, Li, Ou, et al. 'Behavior informatics: A new perspective' IEEE Intelligent Systems, 29 (4) (2014), pp. 62-80. DOI:10.1109/MIS.2014.60

[19]Camacho, S. (Re)reading index cards: the archivist as interpreter in Susan Pui San Lok's *News*, University of Lisbon (2018) http://hdl.handle.net/10451/31668

[20] Carrier, B. File System Forensic Analysis, Boston: Addison Wesley. (2005)

[21] Casey, E. (2019) The chequered past and risky future of digital forensics, Australian Journal of Forensic Sciences, 51:6, 649-664, DOI: 10.1080/00450618.2018.1554090

[22] Casey, E. and Rose, C. W. Chapter 2 - Forensic Analysis, Editor(s): Eoghan Casey, Cory Altheide, Christopher Daywalt, Andrea de Donno, Dario Forte, James O. Holley, Andy Johnston, Ronald van der Knijff, Anthony Kokocinski, Paul H. Luehr, Terrance Maguire, Ryan D. Pittman, Curtis W. Rose, Joseph J. Schwerha, Dave Shaver, Jessica Reust Smith. Handbook of Digital Forensics and Investigation, Academic Press. (2010) pp.21-62, ISBN 9780123742674, DOI:10.1016/B978-0-12-374267-4.00002-1.

[23] Charlton, E., "Working with Legacy Media: A Lone Arranger's First Steps" Practical Technology for Archives, Issue 6 (2016) < https://digitalcommons.ilr.cornell.edu/pta/vol1/iss6/2/>

[24] Chassanoff A., Woods K., Lee C.A. Digital Preservation Metadata Practice for Disk Image Access. In: Dappert A., Guenther R., Peyrard S. (eds) Digital Preservation Metadata for Practitioners. Springer, Cham (2016)

[25] Chassanoff, A. and Altman, M., Curation as "Interoperability With the Future": Preserving Scholarly Research Software in Academic Libraries. Journal of the Association for Information Science and Technology. (2019) DOI:10.1002/asi.24244

[26] Cross, J.E., "Archival reference: state of the art", The Reference Librarian , Vol. 26 No. 56, (1997) pp. 5-25.

[27] Cross, K. & Peck, J., 'Editorial: Special Issue on Photography, Archive and Memory', Photographies, 3:2, (2010) pp.127-138, DOI: 10.1080/17540763.2010.499631

[28]Dekker, A. Collecting and Conserving Net Art. London: Routledge. (2018) DOI:10.4324/9781351208635

[29] Dietrich, D. and Adelstein, F. Archival science, digital forensics, and new media art, Digital Investigation, Volume 14, Supplement 1 (2015) S137-S145, ISSN 1742-2876, DOI:10.1016/j.diin.2015.05.004.

[30] Dobreva, M. and Kim, Y. Automatic metadata generation - use cases. File format metadata (definitive for preservation). (2009) Intrallect. - <https://strathprints.strath.ac.uk/13046/ >

[31] Dodge et al, 'This isn't your father's police force': Digital evidence in sexual assault investigations, Australian & New Zealand Journal of Criminology (2019) DOI:10.1177/0004865819851544

[32] Ermakova, L., Cossu, J.V., Mothe, J., "A survey on evaluation of summarization methods", Information Processing & Management, Volume 56, Issue 5, (2019) pp.1794-1814, ISSN 0306-4573, DOI: 10.1016/j.ipm.2019.04.001

[33] Farmer, D. What are MACtimes?. Dr Dobb's The World of Software Development. (2000) Accessed 25/05/2019. <http://www.drdobbs.com/what-are-mactimes/184404275>

[34] Fürst, G., Ghisletta, P., Lubart. T. The Creative Process in Visual Art: A Longitudinal Multivariate Study. Creativity Research Journal 24:4 (2012) pp.283-295

[35] Forthmann, B., Holling, H., Çelik, P., Storme, M., Lubart. T. Typing Speed as a Confounding Variable and the Measurement of Quality in Divergent Thinking. Creativity Research Journal 29:3 (2017) pp.257-269.

[36] Garfield, R. The Stephen Dwoskin dossier: Introduction, Screen, Volume 57, Issue 1 (2016) pp.71–73, DOI:10.1093/screen/hjw006

[37] Garfinkel, S. Digital media triage with bulk data analysis and bulk_extractor. Computers and Security 32 (2013) pp.56-72. <https://simson.net/clips/academic/2013.COSE.bulk_extractor.pdf>

[38] Given, L. M. and Willson, R. Information technology and the humanities scholar: Documenting digital research practices. Journal of the Association for Information Science and Technology, 69: (2018) pp.807-819. DOI:10.1002/asi.24008

[39] Glăveanu, V. P. Springer briefs in psychology. Distributed creativity: Thinking outside the box of the creative individual. New York, NY, US: Springer Science + Business Media. (2014) DOI:10.1007/978-3-319-05434-6

[40] Gonçalves, M. A., Moreira, B. L., Fox, E. A., Watson, L. T. "What is a good digital library?" – A quality model for digital libraries, Information Processing & Management, Volume 43, Issue 5 (2007) pp.1416-1437. ISSN 0306-4573, DOI:10.1016/j.ipm.2006.11.010.

[41] Goswami, A. Creativity and the quantum: A unified theory of creativity. Creativity Research Journal, 9 (1996) p.47–61.

[42] Greene, M.A., "MPLP: it's not just for processing anymore", The American Archivist , Vol. 73 No. 1, (2010) pp. 175-203.

[43] Grout, H ‘ Archiving critically: exploring the communication of cultural biases’ Spark: UAL Creative Teaching and Learning Journal, Volume 4, Issue 1, (2019) pp.71-75

[44] Guenther, R. Metadata to support long-term preservation of digital assets: PREMIS and its use with METS. In Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop (US-DPIF '10). ACM, New York, NY, USA, , Article 14 (2010) p.6. DOI: 10.1145/2039274.2039288

[45] Guilford, J. P. Creativity.American Psychologist, 5 (1950) p444–454.

[46] Gunn, C., “Putting Personal Digital Archives in Context.” In: The Complete Guide to Personal Digital Archiving. ALA Editions, Chicago, (2018) pp.xi-xxii. ISBN 978-0-8389-1605-6

[47] Hadamard, J. An essay on the psychology of invention in the mathematical field. Princeton, NJ: Princeton University Press (1945)

[48] Hargreaves, C. and Patterson, J. An automated timeline reconstruction approach for digital forensic investigations, Digital Investigation, Volume 9, Supplement, (2012), pp.S69-S79, ISSN 1742-2876, DOI:10.1016/j.diin.2012.05.006.

[49] Harris, V., “How can I help you? Becoming user-centered in special collections”, Archival Issues , Vol. 32 No. 2, (2010) pp. 71-97

[50] Helie, S., Towards a unified neurobiological theory of creative problem solving. The 2013 International Joint Conference on Neural Networks (IJCNN) (2013) pp.1-8.

[51] Hendahewa and Shah, ‘Evaluating user search trails in exploratory search tasks’, Information Processing & Management, Volume 53, Issue 4, (2017), pp.905-922 DOI:10.1016/j.ipm.2017.04.001.

[52] Ho, S. M., Kao, D. and Wu, W.-Y. Following the breadcrumbs: Timestamp pattern identification for cloud forensics, Digital Investigation, Volume 24 (2018) pp79-94, ISSN 1742-2876, DOI:10.1016/j.diin.2017.12.001.

[53] Holm-Hadulla, R. M., The Dialectic of Creativity: A Synthesis of Neurobiological, Psychological, Cultural and Practical Aspects of the Creative Process. Creativity Research Journal 25:3 (2013) pp.293-299.

[54] Hsu, Y. Advanced Understanding of Imagination as the Mediator between Five-Factor Model and Creativity. The Journal of Psychology 153:3 (2019) pp.307-326.

[55] Jatowt, A. et al, “Mapping Temporal Horizons: Analysis of Collective Future and Past related Attention in Twitter”, WWW '15 Proceedings of the 24th International Conference on World Wide Web (2015) pp.484-494, DOI: 10.1145/2736277.2741632

[56] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T.Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM international conference on Multimedia (MM '14). ACM, New York, NY, USA, (2014) pp.675-678. DOI: 10.1145/2647868.2654889

[57] John, J. L., Digital Forensics and Preservation. DPC Technology Watch Report. Digital Preservation Coalition. (2012) SSN: 2048-7916. DOI: 10.7207/twr12-03

[58] Kälber, S., Dewald, A. and Idler, S.. “Forensic Zero-Knowledge Event Reconstruction on Filesystem Metadata.” In: Sicherheit, (2014) pp. 331–343.

[59] Katuu, S., "User studies and user education programmes in archival institutions", Aslib Journal of Information Management, Vol. 67 No. 4, (2015) pp. 442-457. DOI:10.1108/AJIM-01-2015-0005

[60] Kelly, E.J. & Rosenbloom, L. Self Analytics and Personal Digital Archives in University Collections, Collection Management, 44:2-4, (2019) pp.244-258, DOI: 10.1080/01462679.2019.1587672

[61] Kim, Y. "Designated communities": through the lens of the web. International Journal of Digital Curation, 10(1) (2015) pp.184-195. (doi:10.2218/ijdc.v10i1.360)

[62] Kischenbaum, M., Ovenden, R., Redwine, G. Digital Forensics and Born-Digital Content in Cultural Heritage Collections. Council on Library and Information Resources Washington, D.C. With assistance from Rachel Donahue. (2010) https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub149.pdf

[63] Matthew Kirschenbaum, ‘The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary’, Digital Humanities Quarterly, vol. 7, no. 1, 2013, available at <www.digitalhumanities.org/dhq/vol/7/1/000151/000151.html>, accessed 18 April 2019.

[64] Kirschenbaum, M. Track Changes. Harvard University Press. (2016) 378 pages

[65] Kleanthous, S., Christodoulou, D., Papadopoulos, G. A. and Samaras, G., Towards Modelling the User Creative Process in a Sandbox Game. In Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18). ACM, New York, NY, USA, (2018) pp.69-74. DOI: 10.1145/3213586.3226197

[66] Lavoie, B. The Open Archival Information System Reference Model: Introductory Guide. Microform & Imaging Review, 33(2) (2008) pp.68-81. Retrieved 2 Aug. 2019, DOI:10.1515/MFIR.2004.68

[67] Legrand, S., Vanmeert, F., Van der Snickt, G., Alfeld, M., De Nolf, W., Dik, J. and Janssens, K. , Examination of historical paintings by state-of-the-art hyperspectral imaging methods: from scanning infra-red spectroscopy to computed X-ray laminography. Heritage Science20142:13, (2014). DOI:10.1186/2050-7445-2-13

[68] Lienhart, R., Reliable Transition Detection in Videos: A Survey and Practitioner’s Guide. In International Journal of Image and Graphics, 1(3) (2001) pp.469-486.

[69] Lim Rhee, H., “Reflections on Archival User Studies,” Reference & User Services Quarterly 54, no. 4 (Summer 2015): pp.34

[70] Lindley, S., et al., “Rethinking the Web as a Personal Archive” WWW '13 Proceedings of the 22nd international conference on World Wide Web (2013) pp.749-760, DOI: 10.1145/2488388.2488454

[71] Lubart, T.I Models of the Creative Process: Past, Present and Future, Creativity Research Journal, 13:3-4 (2001) pp.295-308, DOI: 10.1207/S15326934CRJ1334_07

[72] Mayer, J., Mutchler, P. and Mitchell, J. C., Evaluating the privacy properties of telephone metadata. Proceedings of the National Academy of Sciences (May 2016), 201508081; DOI: 10.1073/pnas.1508081113

[73] Manoff, M, Archive and Library (2019) DOI: 10.1093/acrefore/9780190201098.013.1017

[74] McCreadie, Rajput, Soboroff, Macdonald, Ounis, 'On enhancing the robustness of timeline summarization test collections', Information Processing & Management, Volume 56, Issue 5, (2019) pp. 1815-1836, DOI: 10.1016/j.ipm.2019.02.006

[75] O'Meara, E. Personal Digital Archiving: DPC Technology Watch Report 15-01. The American Archivist: Spring/Summer 2017, Vol. 80, No. 1 (2017) pp.240-243.

[76] Mele, Bahrainian, and Crestani, 'Event mining and timeliness analysis from heterogeneous news streams', Information Processing & Management, Volume 56, Issue 3, (2019), pp.969-993, DOI:10.1016/j.ipm.2019.02.003.

[77] Minard, A.-L., Speranza, M., Agirre, E., Aldabe, I., van Erp, M., Magnini, B. et al.Semeval, Task 4: Timeline: Cross-document event ordering. Proceedings of the 9th international workshop on semantic evaluation, SemEval '15, Association for Computational Linguistics (2015), pp. 778-786

[78] Nasar, Z. et al. "Textual keyword extraction and summarization: State-of-the-art", Information Processing & Management, Volume 56, Issue 6 (2019) 102088, ISSN 0306-4573, DOI:10.1016/j.ipm.2019.102088.

[79] Ochse, R. Before the gates of excellence: The determinants of creative genius. New York: Cambridge University Press. (1990)

[80] Patrick, C. Creative thought in artists. Journal of Psychology, 4 (1937) pp.35–73.

[81] Plaisant, C., Shneiderman, B., Mushlin, R. An information architecture to support the visualization of personal histories, Information Processing & Management, Volume 34, Issue 5 (1998) pp.581-597. ISSN 0306-4573, DOI:10.1016/S0306-4573(98)00024-7.

[82] Pledge, J., Dickens, E. Process and progress: working with born-digital material in the Wendy Cope Archive at the British Library, Archives and Manuscripts, 46:1 (2018) pp.59-69, DOI: 10.1080/01576895.2017.1408024

[83] Poole, A.H.: Archival divides and foreign countries. historians, archivists, information-seeking, and technology: retrospect and prospect. Am. Archivist **78**(2), (2015) pp.375–433

[84] Post, C. et al, "Digital Curation at Work Modeling Workflows for Digital Archival Materials" in ACM/IEEE Joint Conference on Digital Libraries (2019) DOI: 10.1109/JCDL.2019.00016

[85] Power et al., Improving Archaeologists' Online Archive Experiences Through User-Centred Design, Journal on Computing (2017)

[86] Qazi and Wong, 'An interactive human centered data science approach towards crime pattern analysis', Information Processing & Management, Volume 56, Issue 6, (2019) DOI: 10.1016/j.ipm.2019.102066.

[87] Raghavan, S. & Raghavan, S.V., "Eliciting file relationships using metadata based associations for digital forensics." CSIT (2014) pp.2: 49. DOI:10.1007/s40012-014-0046-4.

[88] Radick, C. "Romance Writers' Use of Archives", Archivaria, Association of Canadian Archivists 81, (2016) pp.45-73,

[89] Rimmer, J., Warwick, C., Blandford, A., Gow, J., Buchanan, G. “An examination of the physical and the digital qualities of humanities research”, Information Processing & Management, Volume 44, Issue 3, (2008) pp.1374-1392. ISSN 0306-4573. DOI:10.1016/j.ipm.2007.09.001.

[90] Rodrigues, Folgado, Belo and Gamboa, ‘SSTS: A syntactic tool for pattern search on time series’, Information Processing & Management, Volume 56, Issue 1, (2019) pp.61-76 DOI:10.1016/j.ipm.2018.09.001.

[91] Sadler-Smith, E. Wallas, “Four-Stage Model of the Creative Process: More Than Meets the Eye?”. Creativity Research Journal 27:4 (2015) pp.342-352.

[92] Sawyer, R. K. “Teaching and Learning How to Create in Schools of Art and Design”. Journal of the Learning Sciences 27:1 (2018) pp137-181.

[93] Seadle, M.: Managing and mining historical research data. Lib. Hi Tech. **34**(1), (2016) pp.172–179

[94] Schinas, M., Papadopoulos, S., Kompatsiaris, Y. and Mitkas, P. A., Event Detection and Retrieval on Social Media, CoRR,abs/1807.03675 (2018). <http://arxiv.org/abs/1807.03675>

[95] Schneier, B., Data And Goliath : the Hidden Battles to Collect Your Data and Control Your World. New York, N.Y. :W.W. Norton & Company (2015).

[96]Sinn, D., Kim, S. and Syn, S., "Personal digital archiving: influencing factors and challenges to practices", Library Hi Tech, Vol. 35 No. 2, (2017) pp. 222-239. DOI:10.1108/LHT-09-2016-0103

[97]Stein, M. I., Stimulating creativity: Individual procedures. New York: Academic. (1974)

[98] Sunde, N., Dror, I.E. “Cognitive and human factors in digital forensics: Problems, challenges, and the way forward”, Digital Investigation, Volume 29, (2019) pp.101-108 DOI: 10.1016/j.diin.2019.03.011

[99] Taramigkou, Apostolou, and Mentzas, ‘Leveraging exploratory search with personality traits and interactional context’, Information Processing & Management, Volume 54, Issue 4 (2018), pp.609-629, DOI:10.1016/j.ipm.2018.04.001.

[100] Taylor, I. A. “The nature of the creative process.” in P. Smith(Ed.), Creativity: An examination of the creative process (1959) pp.51–82. New York: Hastings House.

[101] Turnbull and Wheeler. “The advertising creative process: A study of UK agencies.” Journal of Marketing Communications 23:2 (2017) pp.176-194.

[102] Ulger, K. Comparing the effects of art education and science education on creative thinking in high school students. Arts Education Policy Review 120:2 (2019) pp.57-79.

[103] Wallas, G. The art of thought. New York: Harcourt Brace. (1926)

[104] Vilar et al., ‘Information Competencies of Historians as Archive Users: A Slovenia/UK Comparison’, CCIS, volume 676 (2017) pp.519-529

[105] Witten, I. H., Bainbridge, D. and Nichols, D. M. “Chapter 5 - Multimedia: More raw material”, in eds. Ian H. Witten, David Bainbridge, David M. Nichols, In The Morgan Kaufmann Series in Multimedia Information and Systems, How to Build a Digital Library (Second Edition), Morgan Kaufmann, (2010) pp.215-284, ISBN 9780123748577, DOI:10.1016/B978-0-12-374857-7.00005-0.

[106] Woods, K. Lee, C. A. and Garfinkel, S. “Extending digital repository architectures to support disk image preservation and access.” in Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (JCDL '11). ACM, New York, NY, USA (2011) pp. 57-66. DOI:10.1145/1998076.1998088

[107] Yuelin Li, Chang Liu, “Information Resource, Interface, and Tasks as User Interaction Components for Digital Library Evaluation”, Information Processing & Management, Volume 56, Issue 3, (2019), pp.704-720, ISSN 0306-4573, DOI:10.1016/j.ipm.2018.10.012.

[108] Zhao, Y. et al, “Abnormal Activity Detection Using Spatio-Temporal Feature and Laplacian Sparse Representation” ICONIP 2015: Neural Information Processing (2015) pp 410-418 DOI: 10.1007/978-3-319-26561-2_49